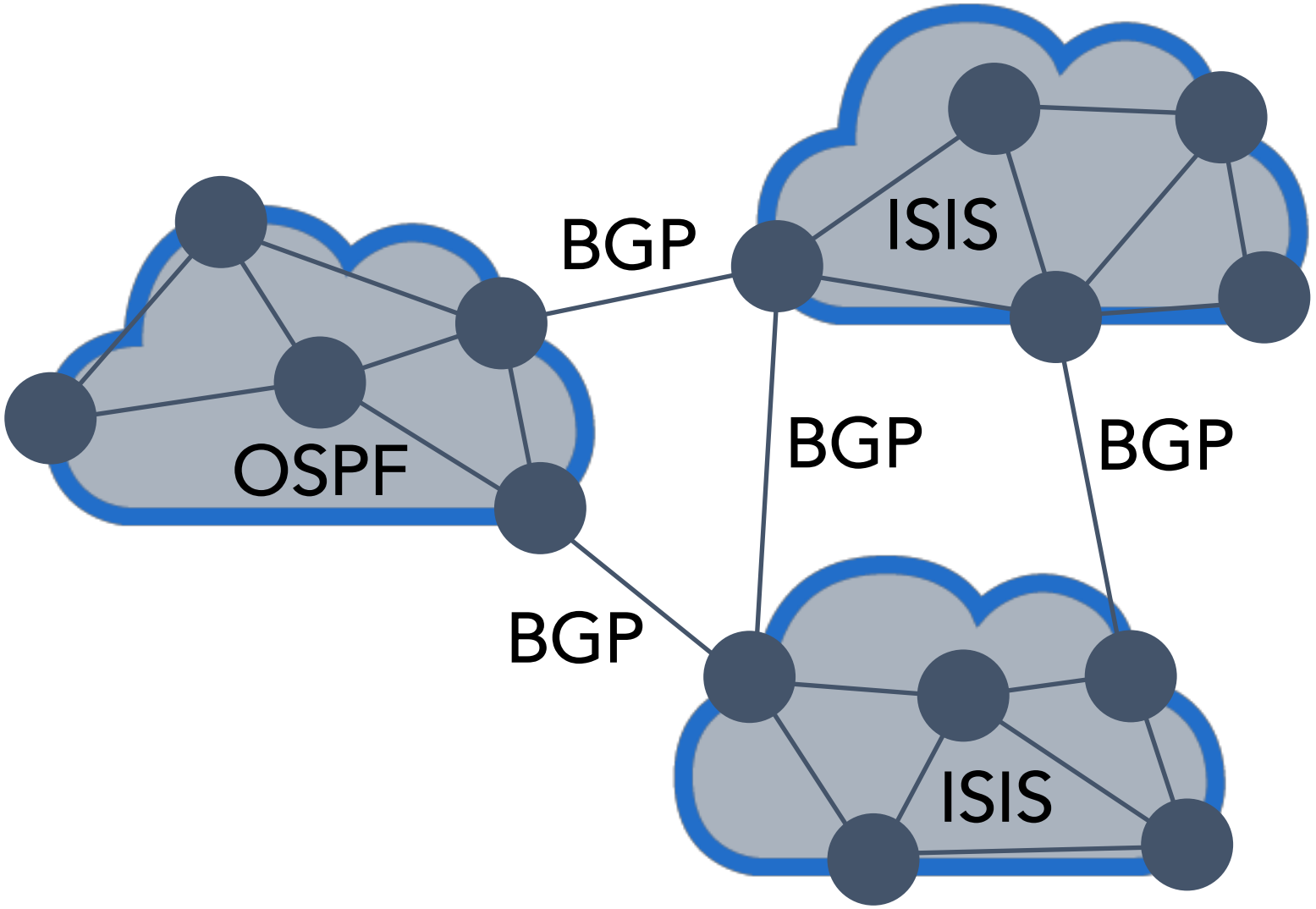
Routing in the Classical Internet

QuTech, Delft

Intra-domain vs inter-domain routing

- **Intra-domain routing protocols (IGP = Interior Gateway Protocol)**
  - Exchange routes *inside* autonomous system
  - All routers owned and operated by same organization
  - Higher level of trust, more information available
  - Emphasis on shortest path
  - Example: OSPF, ISIS, EIGRP, RIP, RIFT, BGP (!), ….
- **Inter-domain routing protocols (EGP = Exterior Gateway Protocol)**
  - Exchange routes *between* autonomous systems
  - Routers owners by different organizations
  - Lower level of trust, less information available, real money is involved
  - Emphasis on implementing business policies (customers, transit, peering, …)
  - Only one protocol: BGP version 4

Intra-domain vs inter-domain routing



Routing protocol algorithms

- **Link State Routing**
  - Every router discovers full topology and computes shortest path
  - The dominant algorithm for intra–domain routing
  - Open Shortest Path First (OSPF)
  - Intermediate-System to Intermediate-System (ISIS)
- **Vector Routing**
  - Distance-Vector or Path-Vector
  - Routers do not know full topology
  - Each router locally choses best path and propagates that path
  - Routing Information Protocol (RIP)
  - Border Gateway Protocol (BGP)

Unicast vs multicast routing

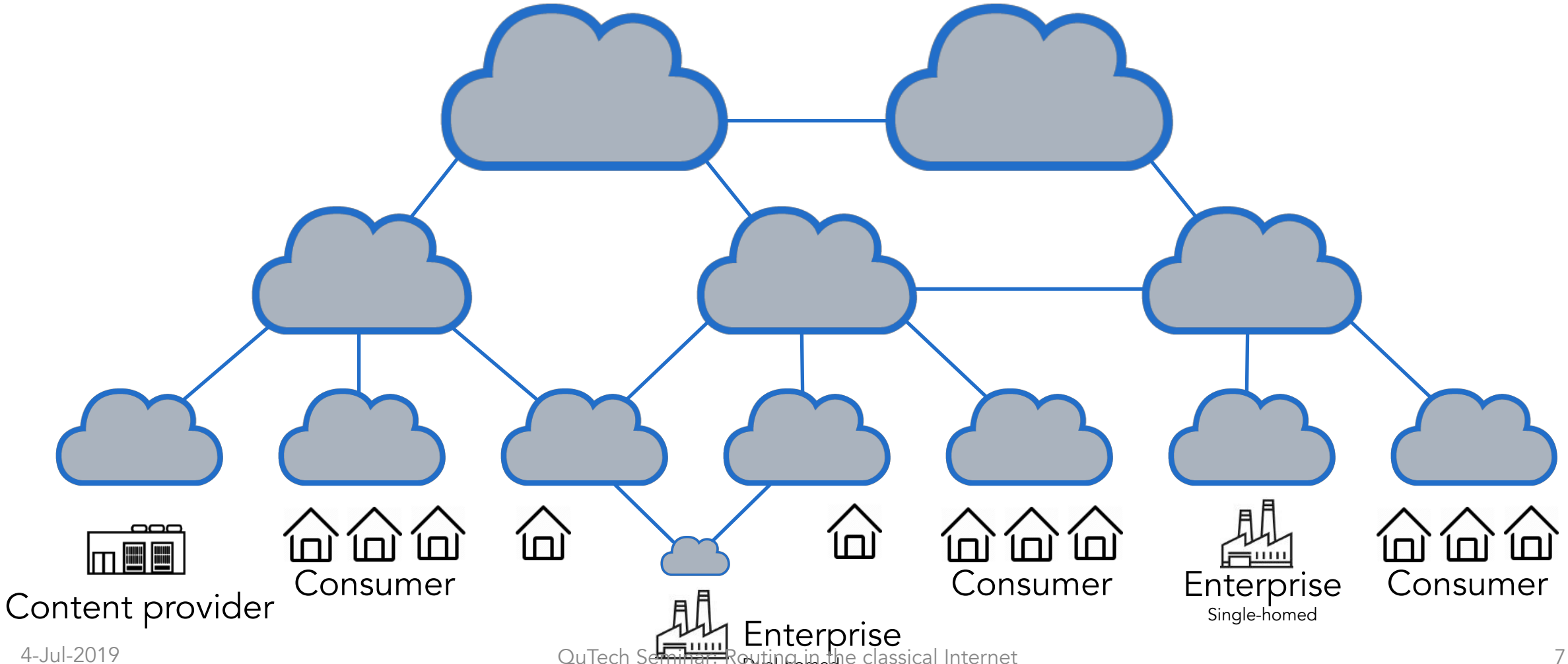
- Unicast routing protocols
  - Each packet is delivered to single destination
  - Destination IP address is unicast IP address (could be anycast)
  - Most traffic on the internet (including video-on-demand)
  - Most routing protocols: OSPF, ISIS, BGP, …
- Multicast routing protocols
  - Each packet is delivered to a group of zero or more destinations (“listeners”)
  - Destination IP address is multicast IP address (224.x.x.x – 239.x.x.x)
  - Example application: live broadcast TV
  - Signaling protocols for listeners to join and leave a group: IGMP, MLD
  - Routing protocols to create distribution tree in the network: PIM
  - Not relevant for Quantum networks (no-cloning)? Or maybe multi-partite entanglement?

Deep dive into BGP

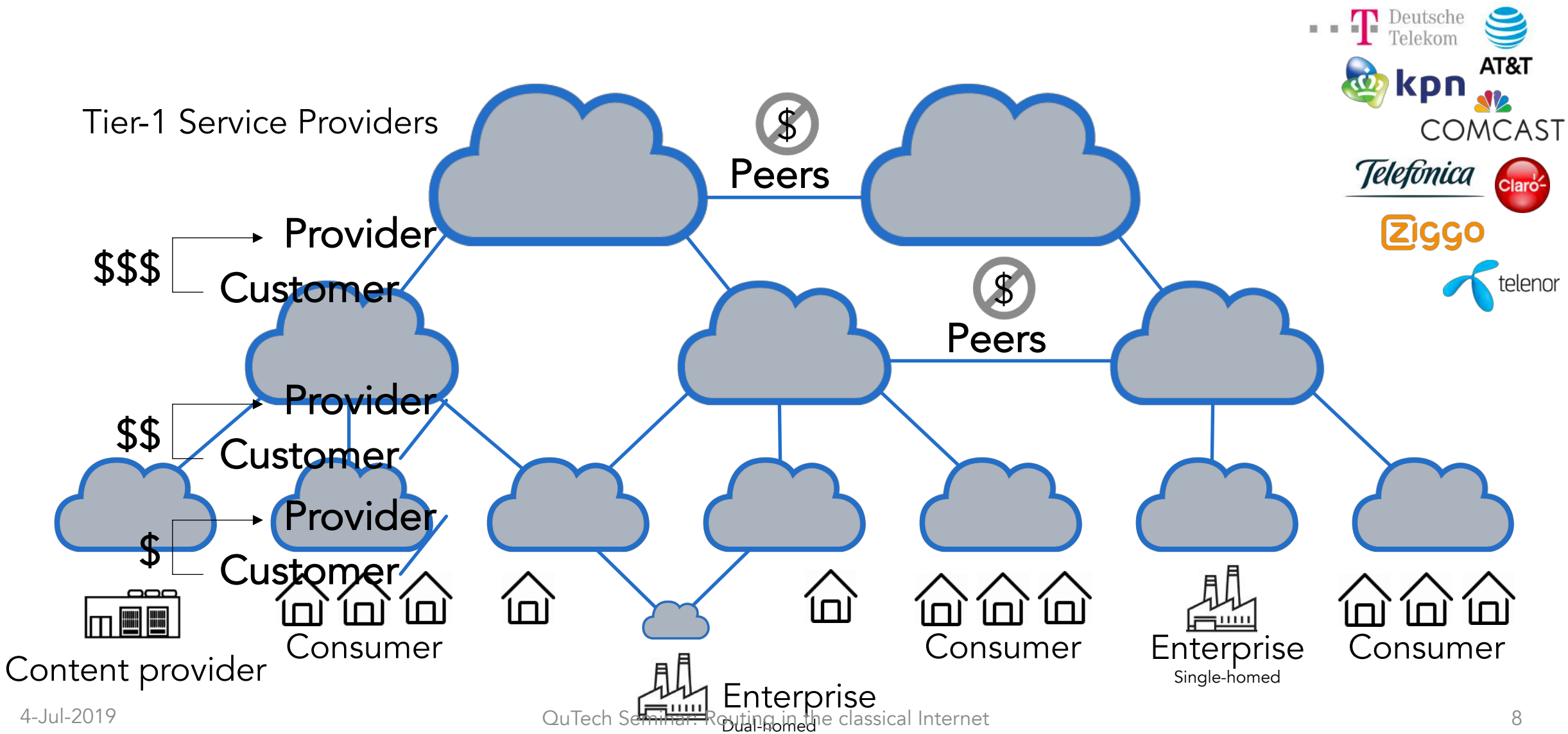
What makes BGP so interesting?

- It is the one and only routing protocol that is **used across the entire Internet**; to understand the Internet you must understand BGP.
- It is not just about finding the shortest path; it has a very rich **"policy" framework** to reflect business requirements.
- The **scaling** challenges are enormous; store 1+ million routes in the route table; advertise 1+ billion routes in a few minutes; …
- The **feature** richness is bewildering:
  - Many services beyond IPv4/IPv6 Internet: L3VPN, VPLS, EVPN, LS-BGP, …
  - Cannot upgrade the Internet all at once: capabilities, etc.
  - There is no scheduled downtime for the Internet: NSF, NSR, …

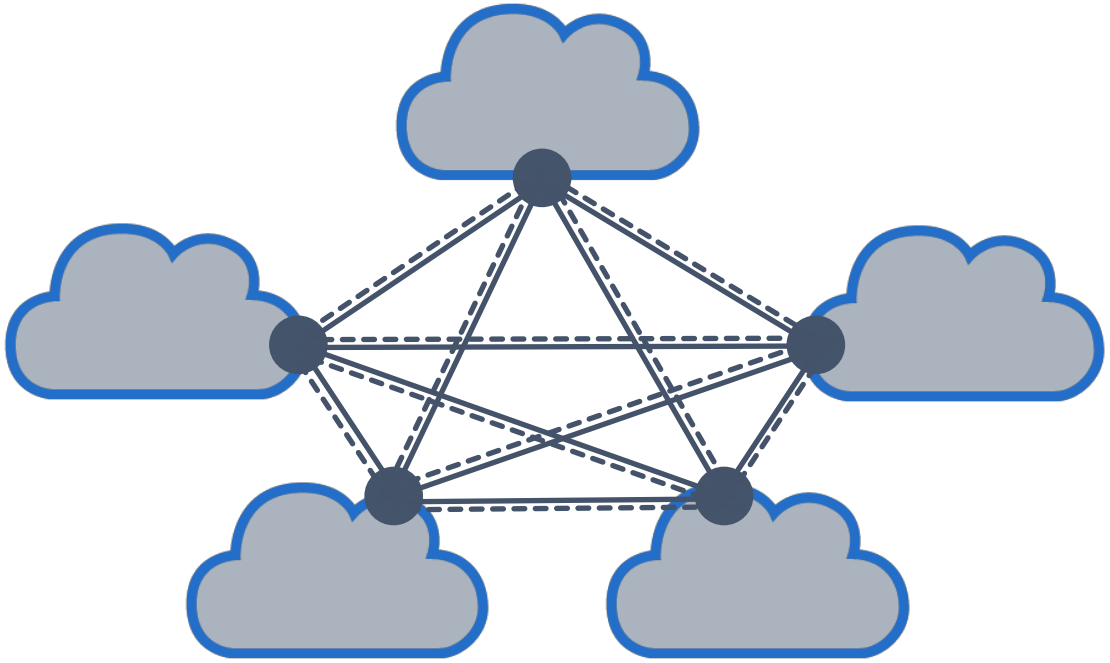
Example Internet topology



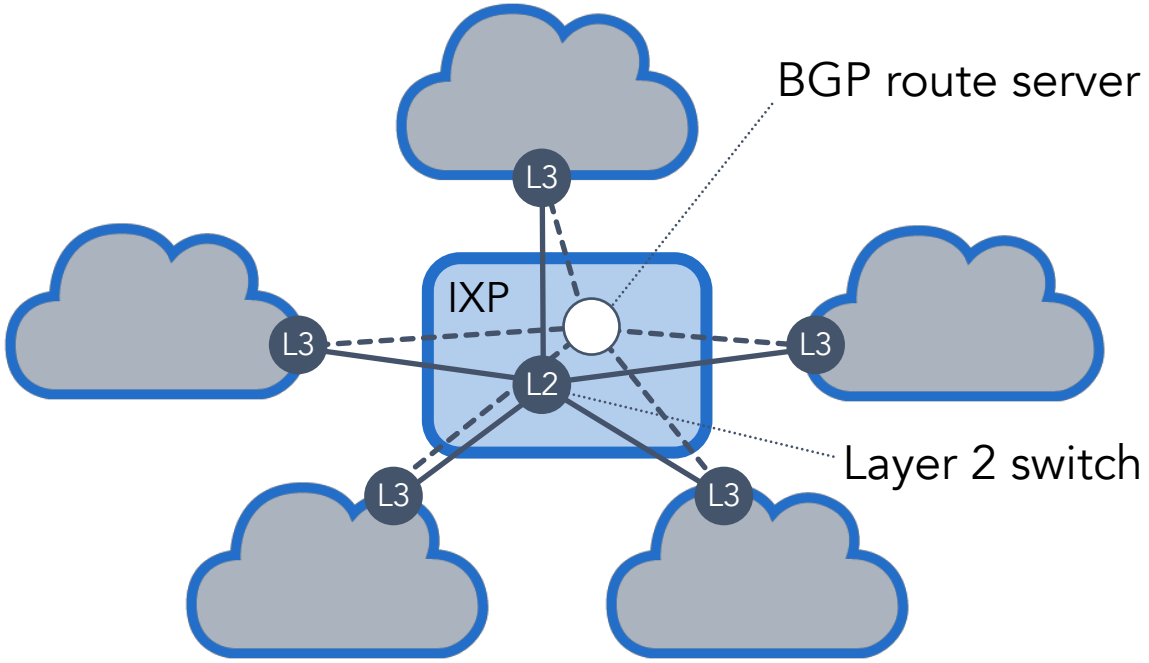
Commercial relationships in the Internet



Internet eXchange Point (IXP)



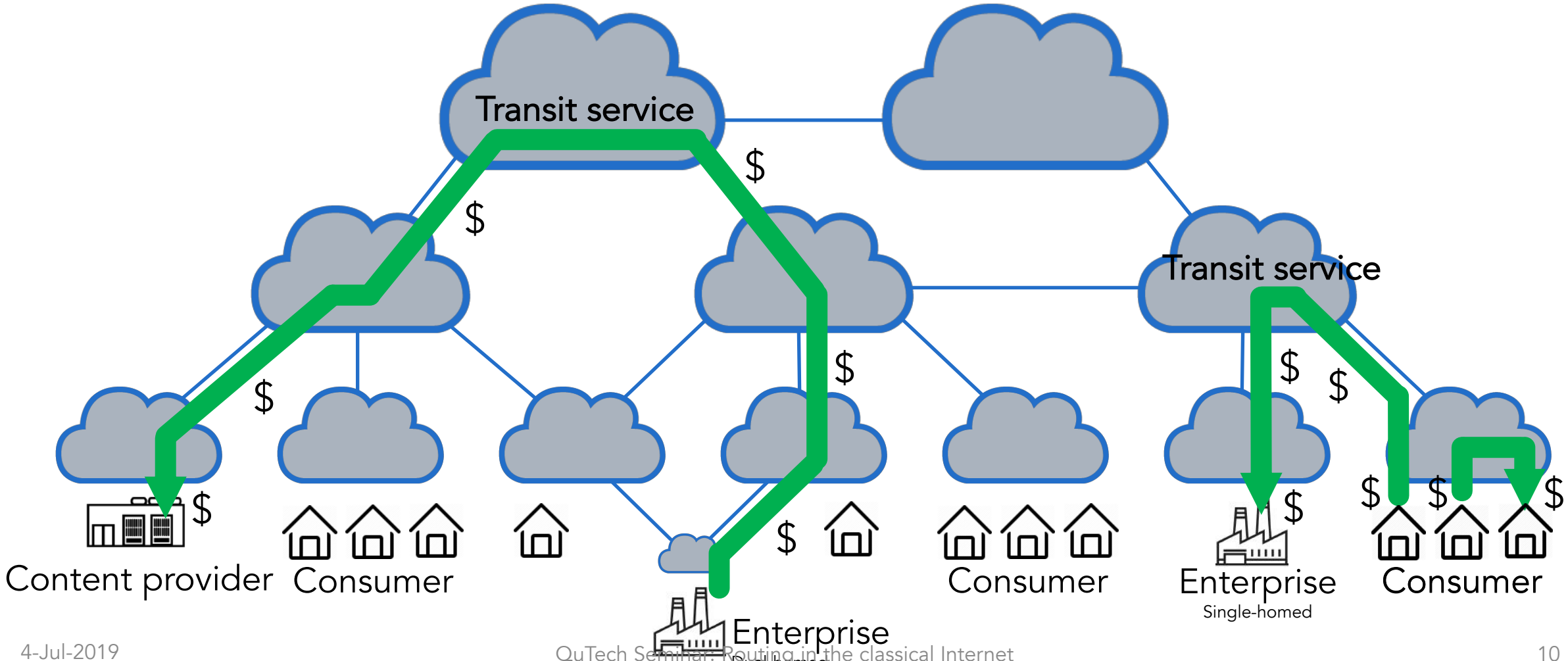
Bilateral peering



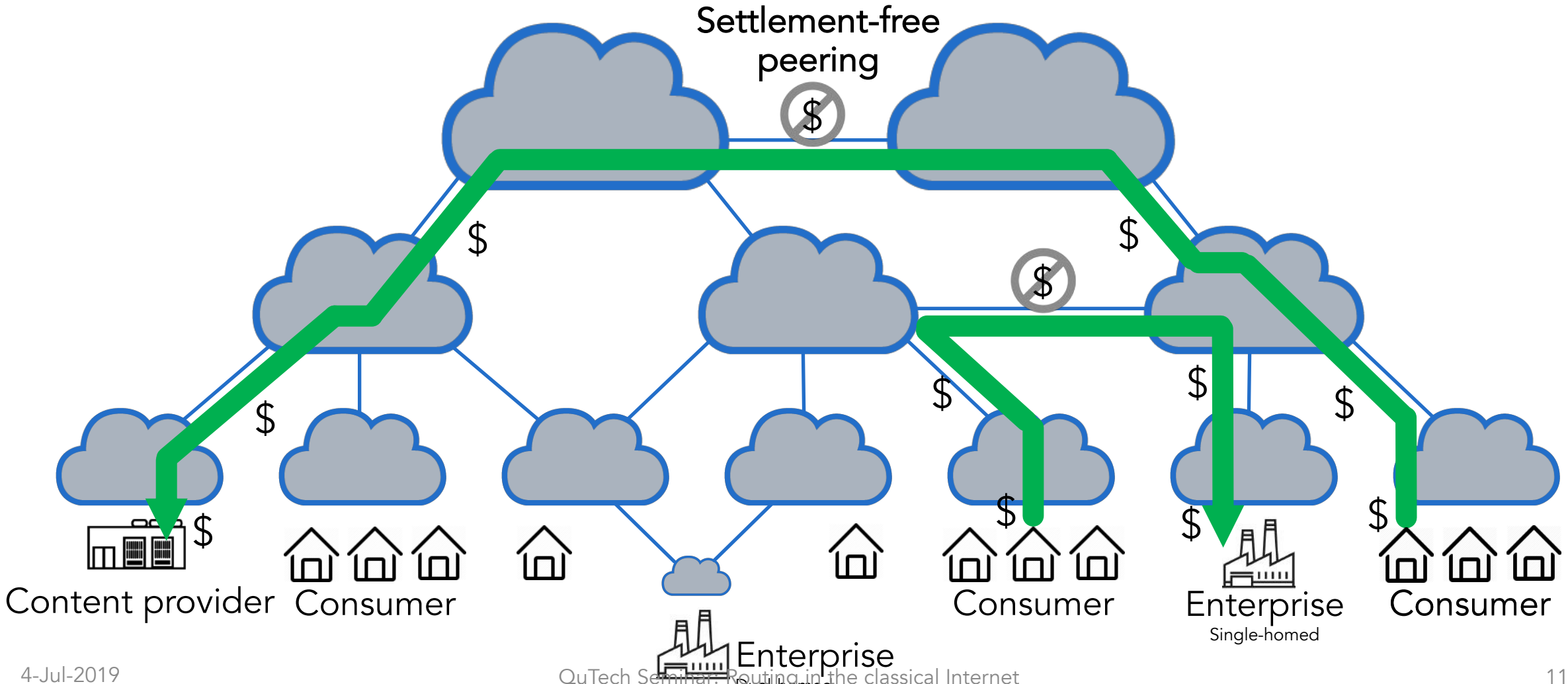
Internet Exchange Point

amsix linx NLix The Interconnect Exchange

Provider – customer traffic

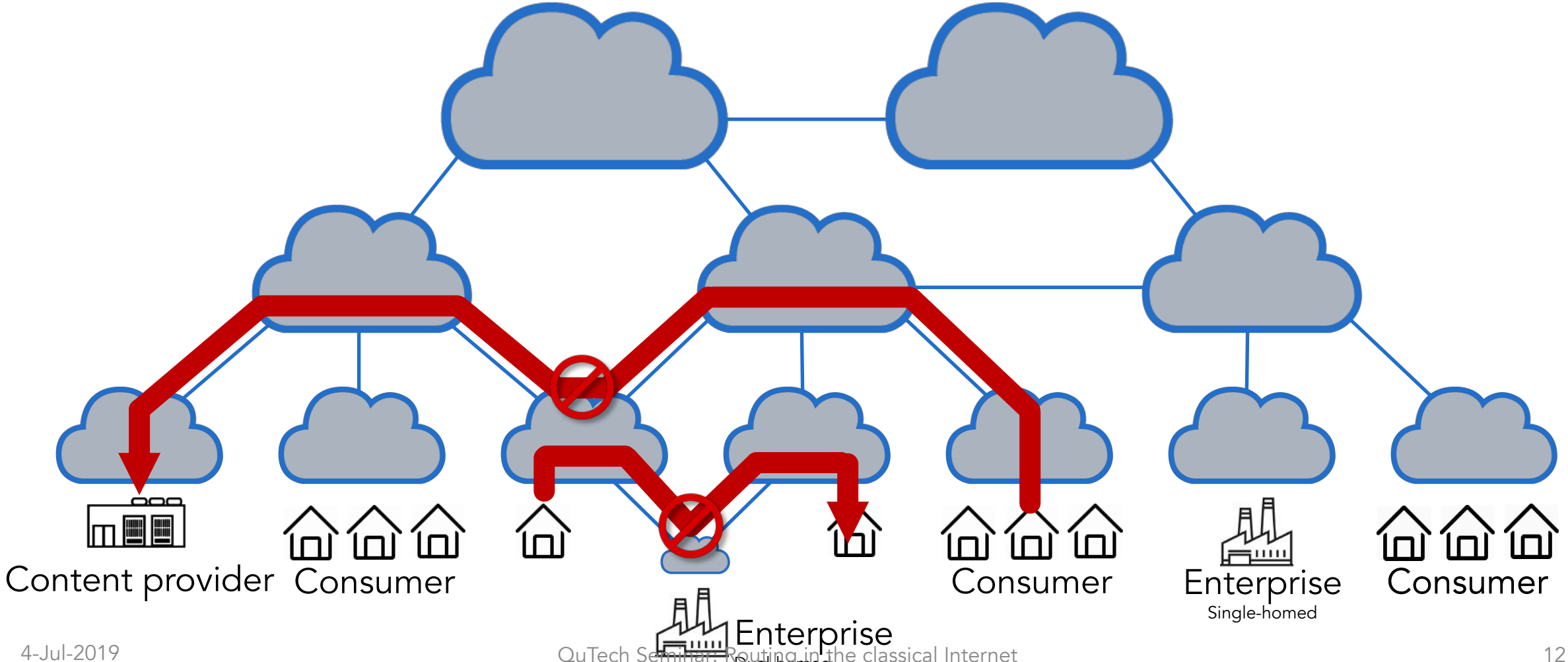


Peer – peer traffic



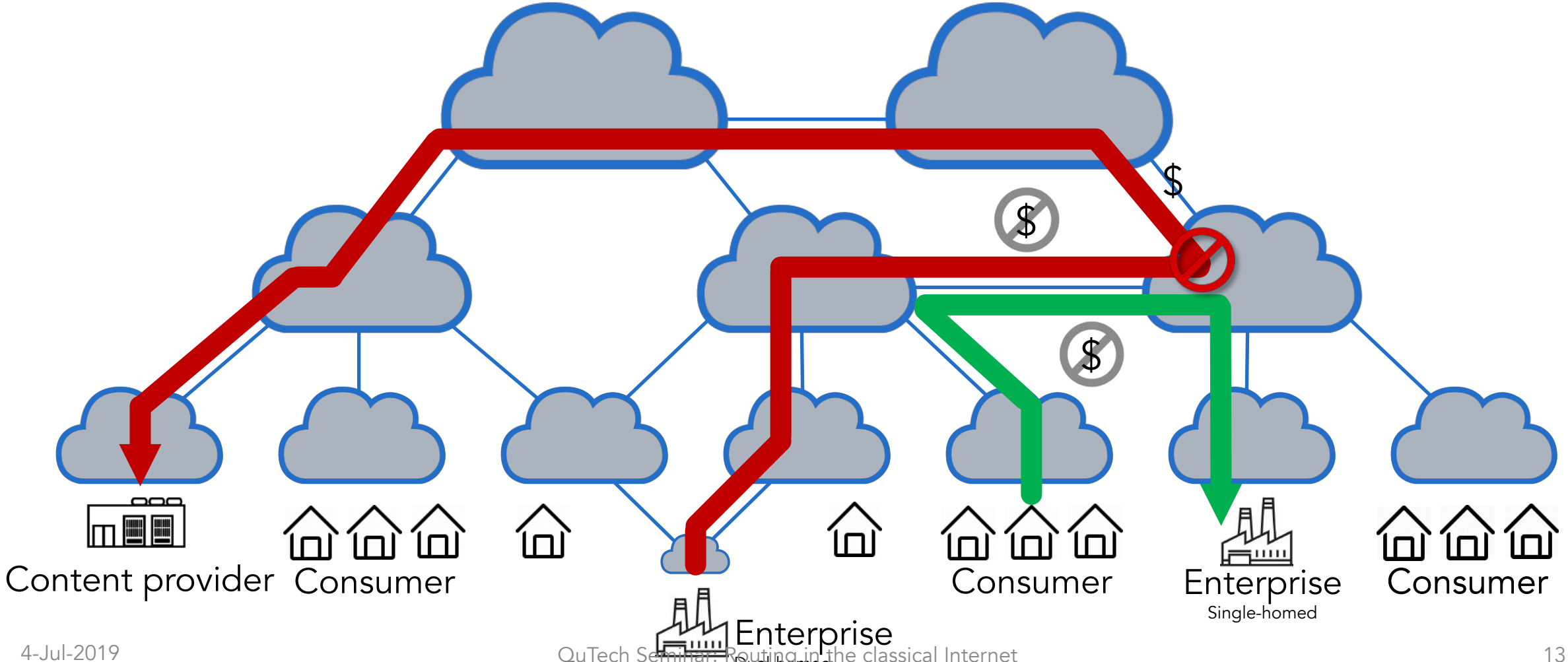
Valley-free routing

Providers must not transit traffic through their customers

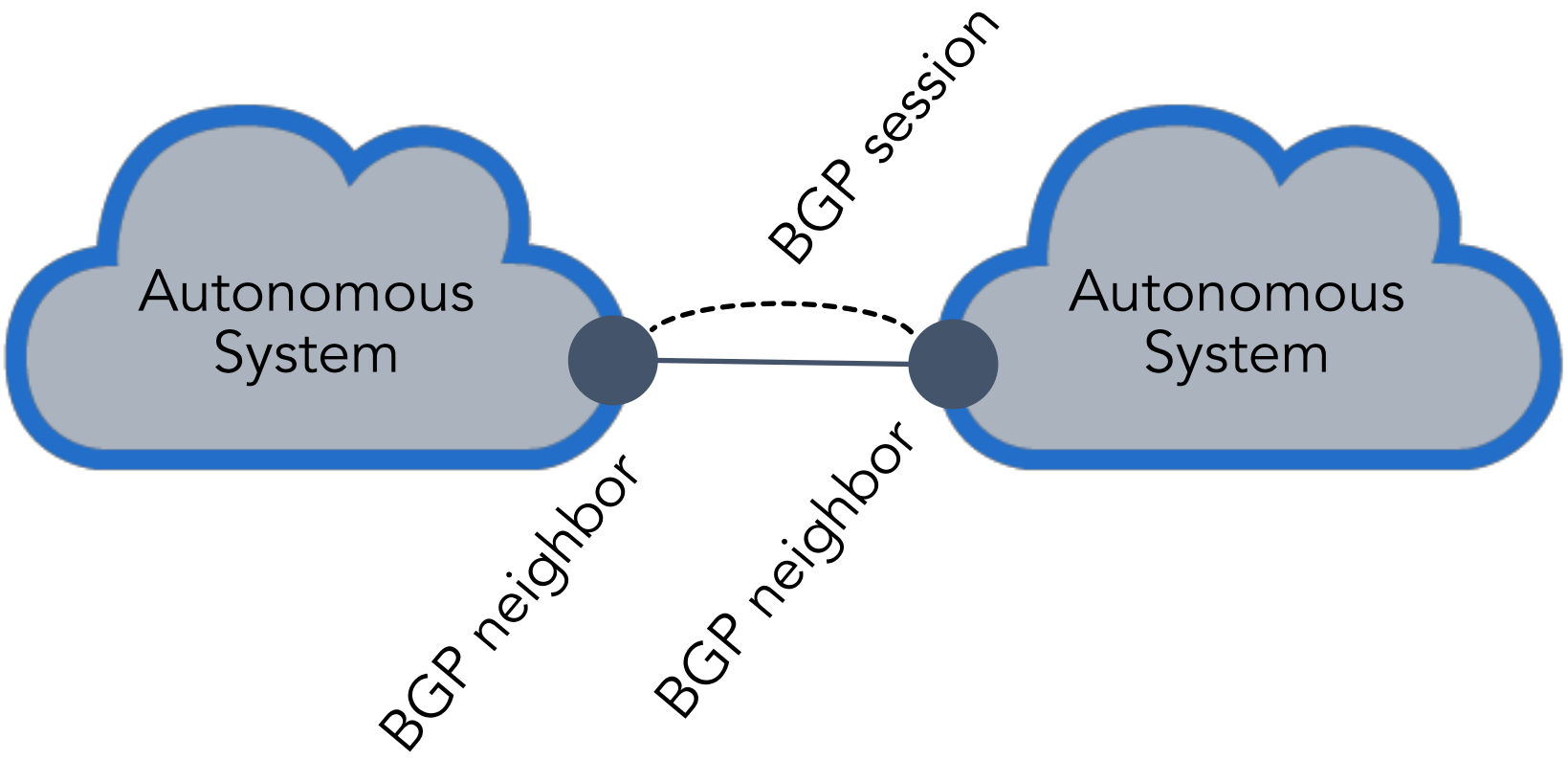


Peering is for direct customer traffic only

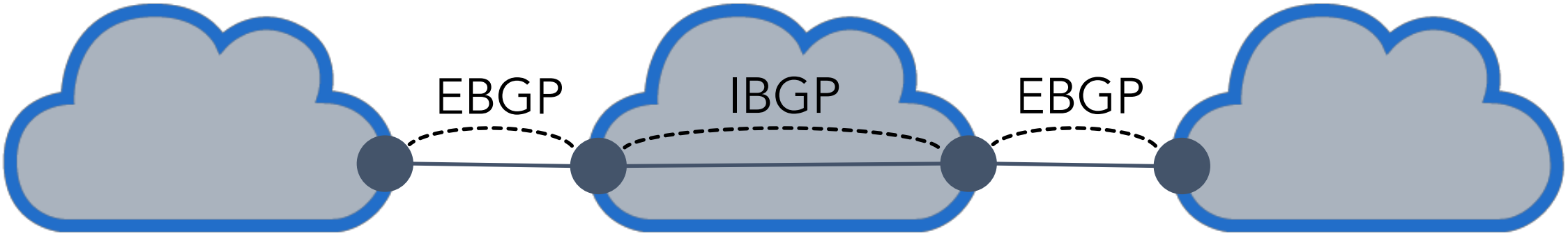
Don’t abuse peering to make someone else pay for transit



BGP neighbors and BGP sessions



IBGP versus EBGP



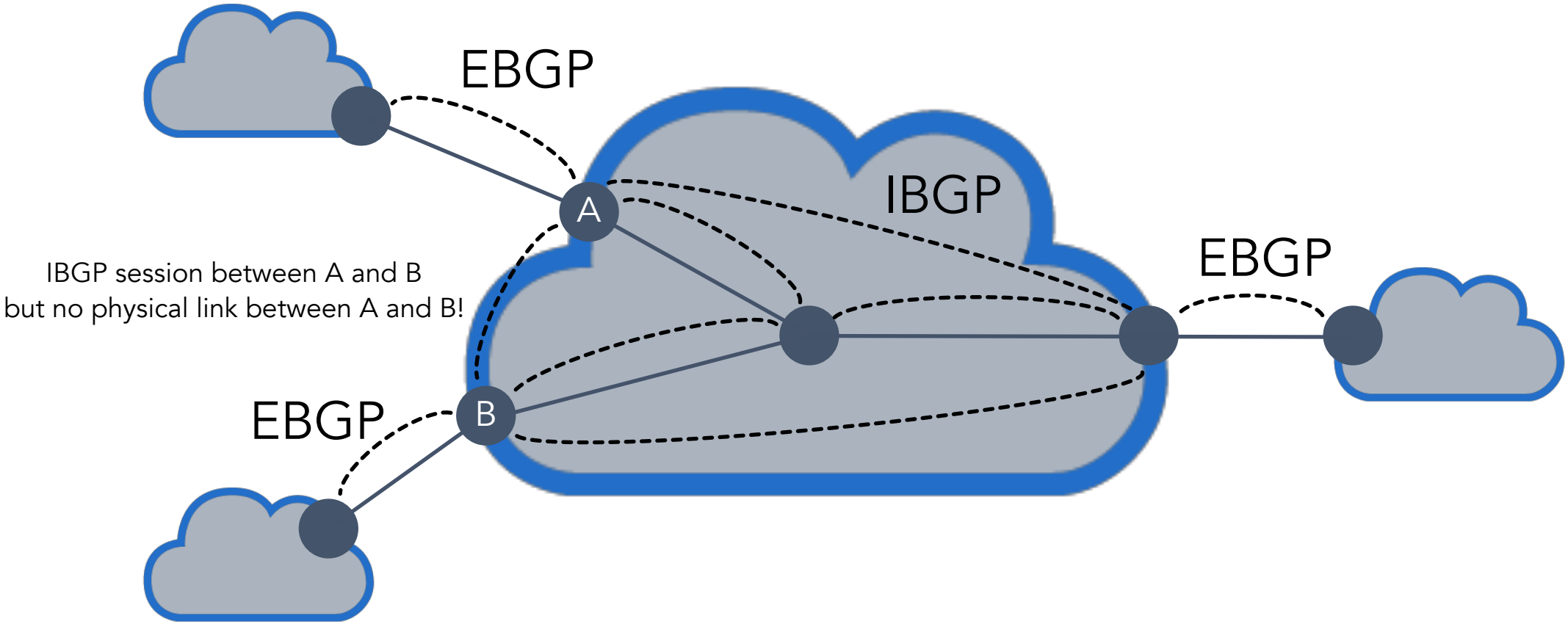
EBGP = External BGP = Between different AS

Normally between directly connected EBGP neighbors

IBGP = Internal BGP= Inside one AS

Often, IBGP neighbors are not directly connected (multi-hop BGP session)

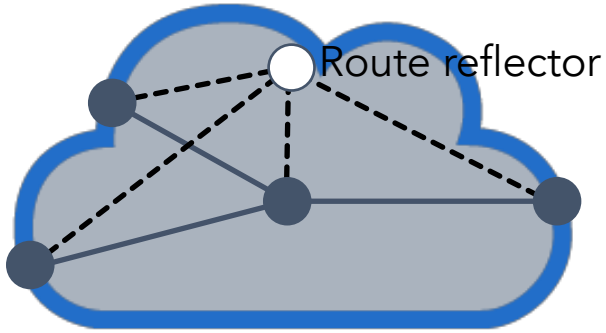
IBGP full mesh



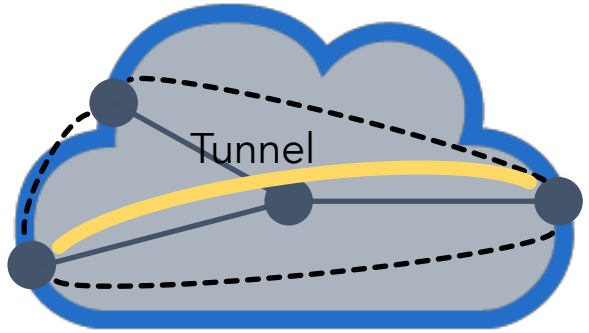
Full mesh of N^2 IBGP multi-hop IBGP sessions

Because no IBGP to IBGP route propagation, because AS-Path not used for IBGP loop detection

Ways to reduce IBGP full mesh

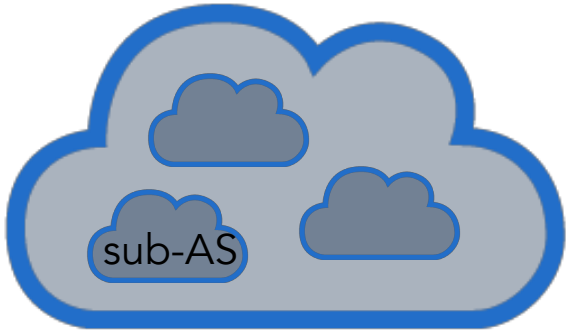


Route reflector



BGP-free core

Tunnel mesh (LSP or GRE)



Confederation

"Nested" AS

BGP sessions run over TCP

- BGP relies on TCP for reliability and flow-control
- BGP is not a periodic protocol (advertise once, withdraw explicitly)
- TCP cannot detect failure in the absence of traffic (hence KEEPALIVE)
- Chicken-and-egg problem? No, TCP uses IGP routes to connect.

OPEN: Establishing a BGP session

- OPEN message is sent to initialize BGP session
- First message after TCP connection established
- Both sides initiate TCP connection; rules for collisions
- Fields:
  - Version: 4
  - Local AS number
  - BGP identifier: router ID
  - Proposed hold time: time-out for KEEPALIVE
  - Capabilities: negotiate optional features (multiprotocol, route refresh, 4-octet AS numbers, many more…)

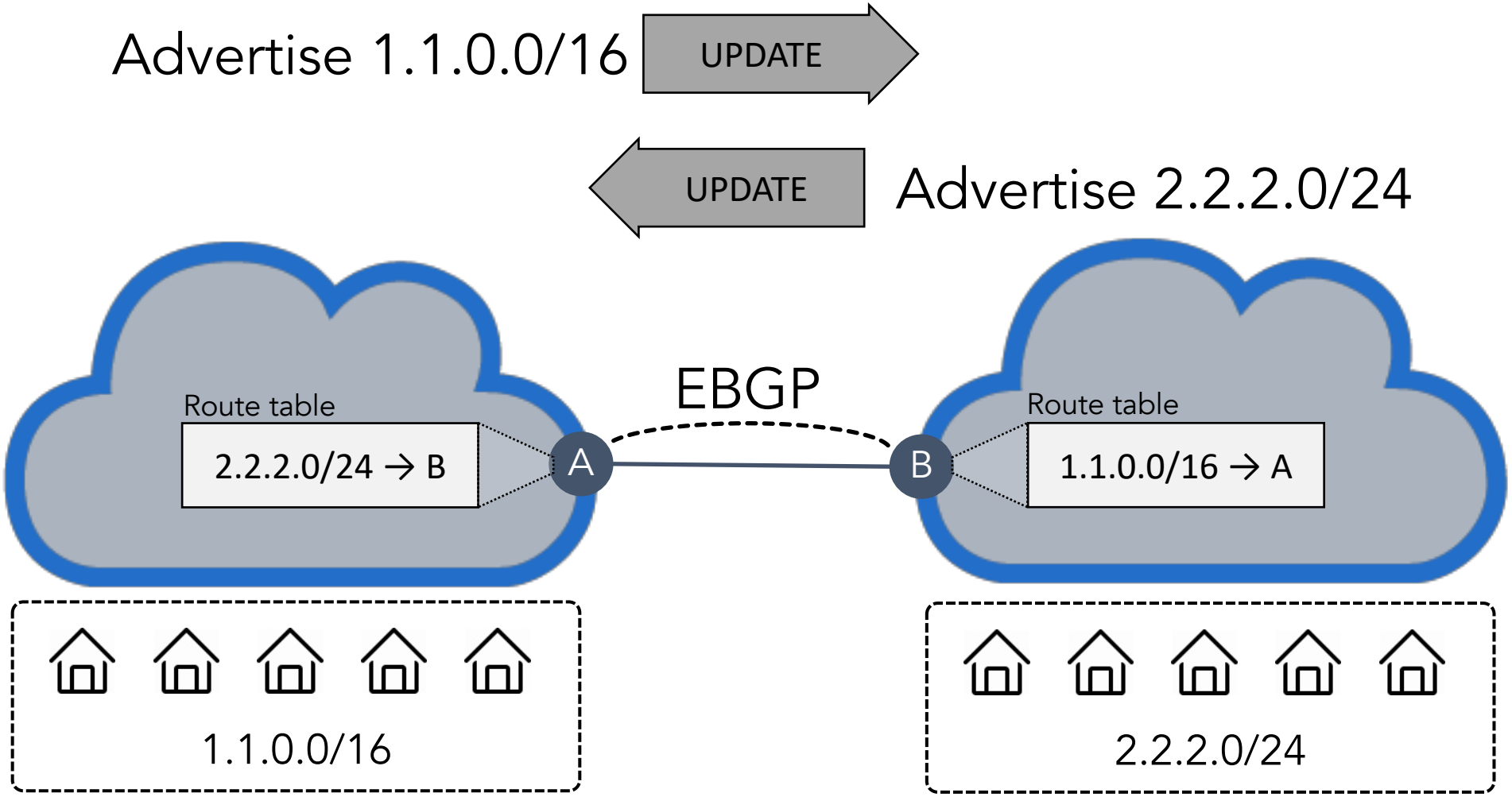
NOTIFICATION: Closing a BGP session

- NOTIFICATION message is sent to terminate BGP session
- Scheduled ("Cease") or error condition
- TCP connection is closed
- All routes from BGP neighbor are removed when BGP session goes down for any reason (exception: graceful restart).
- This is a "heavy hammer"
- Fields:
  - Error code: general category of problem
  - Error subcode: more detailed cause of error
  - Error data: more details about the error

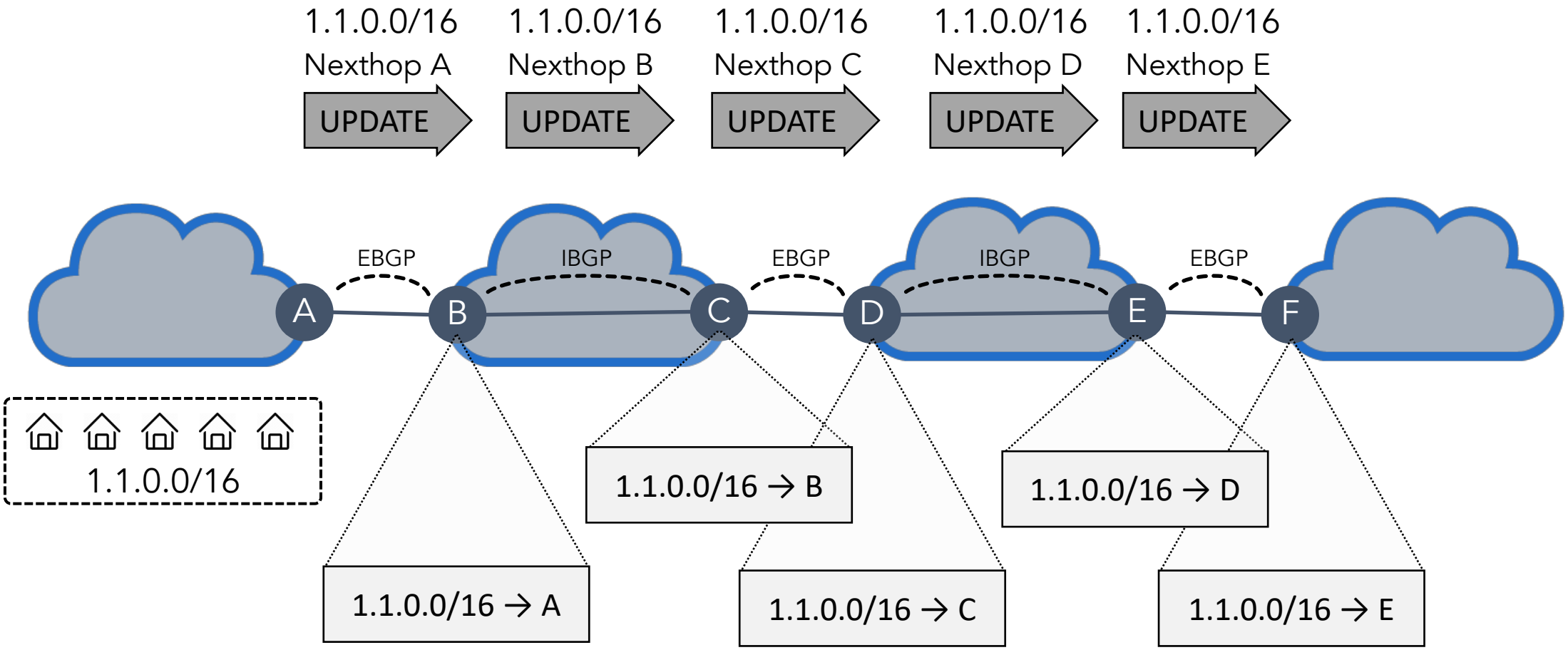
KEEPALIVE: Check liveness

- KEEPALIVE is sent periodically to check liveness of TCP connection in absence of other traffic
- Sent 1/3 of negotiated hold-time (typically every 30 seconds, but often configured to be sent much faster)
- If not received 3x in a row, hold time expired NOTIFICATION is sent.

UPDATE: Advertise routes

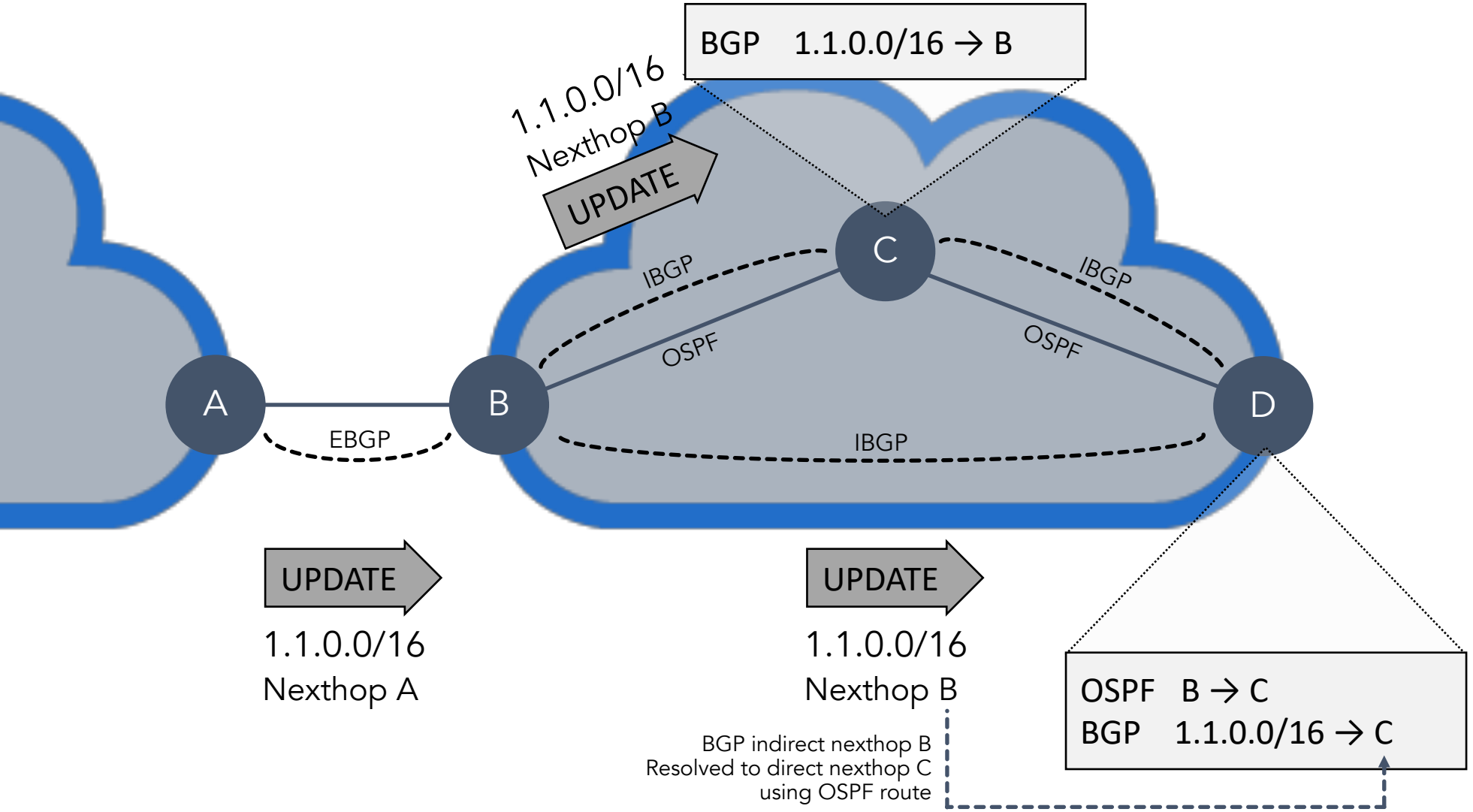


UPDATE propagation over multiple ASs

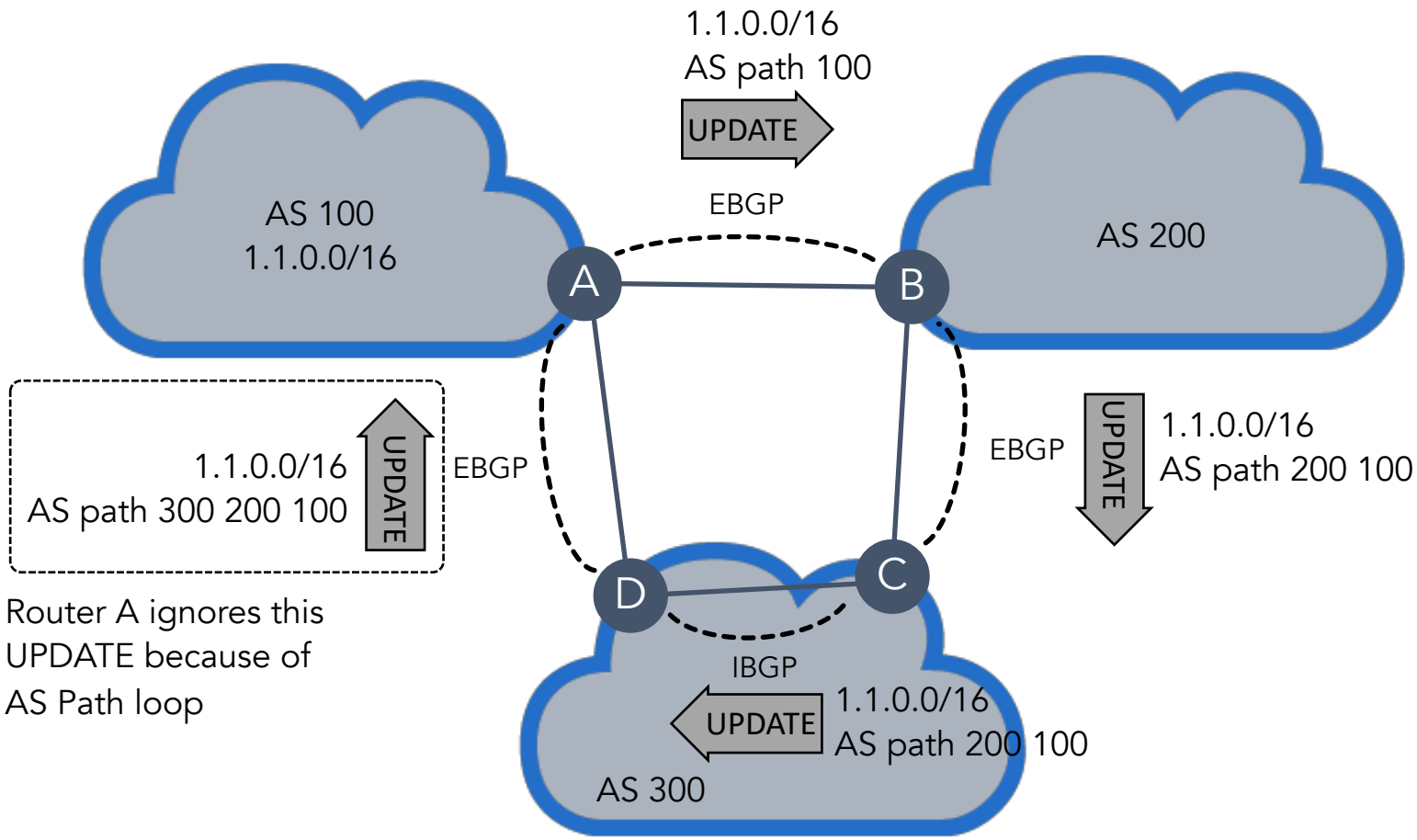


Note: for the sake of simplicity example assumes single-hop IBGP sessions and next-hop-self on IBGP sessions

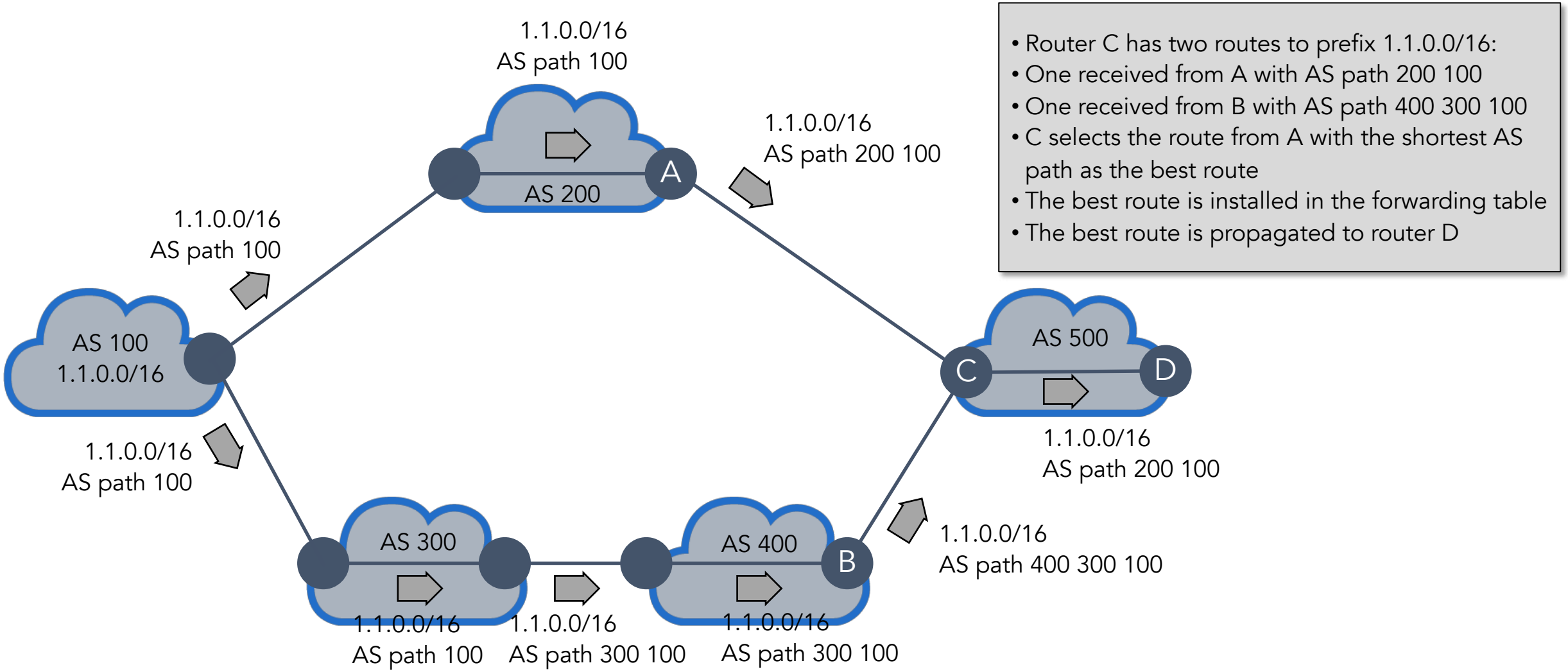
Indirect nexthop resolution



AS path for loop detection



AS path for shortest path selection



- Router C has two routes to prefix 1.1.0.0/16:
- One received from A with AS path 200 100
- One received from B with AS path 400 300 100
- C selects the route from A with the shortest AS path as the best route
- The best route is installed in the forwarding table
- The best route is propagated to router D

Information in a BGP UPDATE

- Withdrawn prefixes (NLRI)
- Advertised prefixes (NLRI)
- Attributes of advertised prefixes
  - Origin
  - AS Path: loop detection and shortest path selection
  - Nexthop: indirect nexthop, which is resolved to direct nexthop using IGP
  - Multi-Exit Discriminator (MED): indicate preferred entry-point into my AS
  - Local preference: indicated preferred exit-point from my AS
  - Atomic aggregate
  - Aggregator: who created an aggregate route (summary route)
  - Optional extensions to BGP added ~30 more attributes, e.g. cluster-list
    https://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-2

Network Layer Reachability Information (NLRI)

- Originally IPv4 prefix
- Later generalized to other address families, e.g. IPv6
- Later generalized to key of table being synchronized
  - Layer 3 Virtual Private Networks (L3 VPN)
  - Ethernet Virtual Private Network (EVPN)
  - Etc.
- BGP has *almost* turned into an eventually consistent general-purpose database synchronization protocol (controversial)

The decision process (= path selection)

1. Nexthop must be reachable
2. Lowest weight
3. Highest local preference
4. Prefer locally generated
5. Shortest AS path length
6. Origin: prefer
7. Lowest MED
8. Prefer EBGP over IBGP
9. Lowest metric of IGP route that resolves indirect nexthop
10. Declare ECMP if multipath is enabled
11. For external paths, prefer first received
12. Lowest router ID
13. Lowest cluster list
14. Lowest neighbor address

Example BGP route (Juniper)

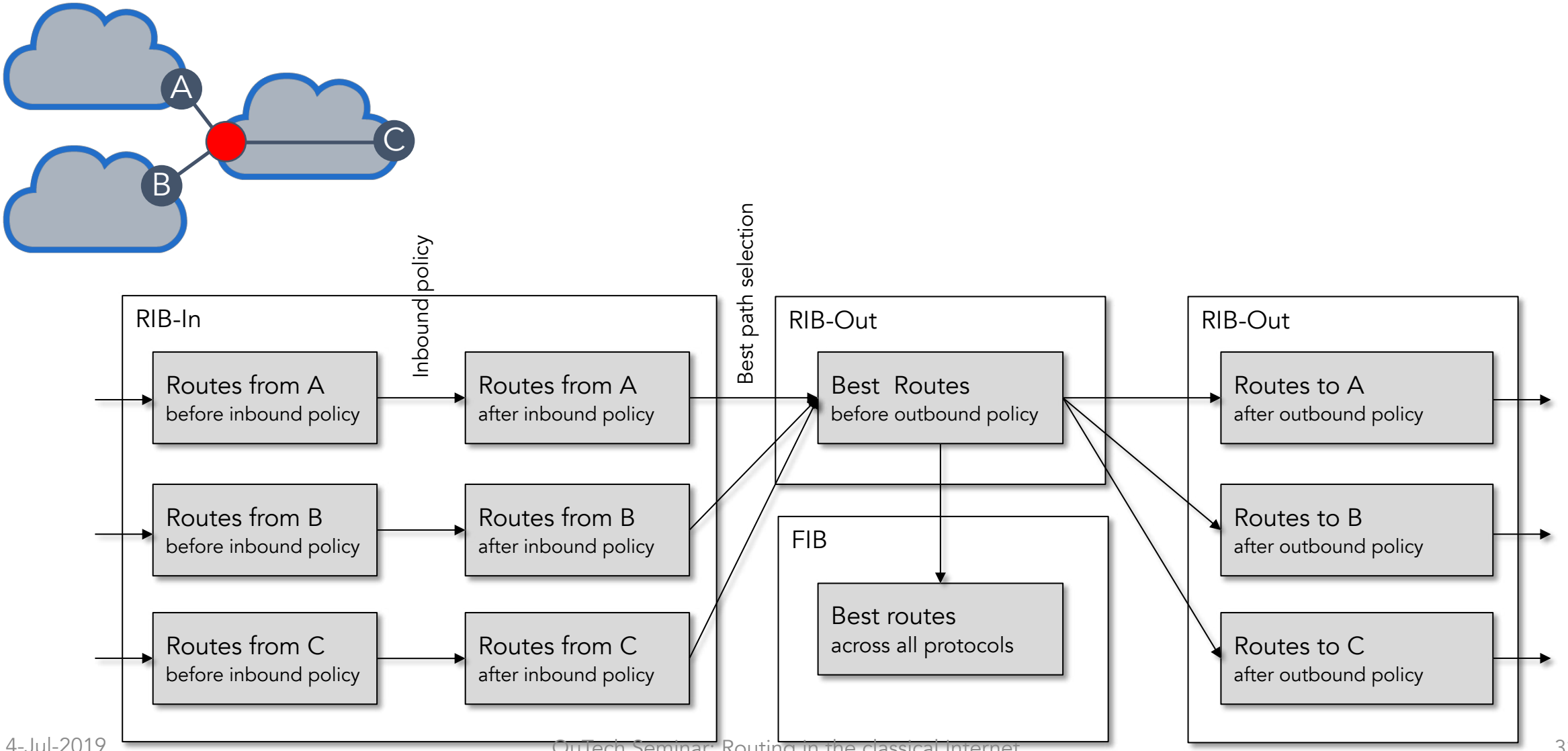
```
user@R4> show route 100.100.1.0 detail
inet.0: 20 destinations, 24 routes (20 active, 0 holddown, 0 hidden)

100.100.1.0/24 (2 entries, 1 announced)
         *BGP    Preference: 170/-201
                 Source: 10.0.0.2
                Next hop: 10.1.24.1 via so-0/0/3.0, selected
                Protocol next hop: 10.0.0.2 Indirect next hop: 8644000 277
                State: <Active Int Ext>
                Local AS: 65002 Peer AS: 65002
                Age: 2:22:34    Metric: 5       Metric2: 10
                Task: BGP_65002.10.0.0.2+179
                Announcement bits (3): 0-KRT 3-BGP.0.0.0.0+179 4-Resolve inet.0
                 AS path: 65001 I
                 Localpref: 200
                Router ID: 10.0.0.2
         BGP    Preference: 170/-101
                 Source: 10.1.45.2
                [...]
```

BGP policies

- Policy is *the* central concept in BGP
  - It allows BGP to implement reflect business rules (money is involved)
  - BGP policy is *much* richer than policy in any other routing protocol
- Policy rules describe:
  - Which routes should be accepted from a BGP neighbor
  - Which routes should be advertised to a BGP neighbor
  - How the attributes of accepted and advertised routes should be modified
- It explains why BGP is a "path vector" protocol
- Each router vendor has its own policy language

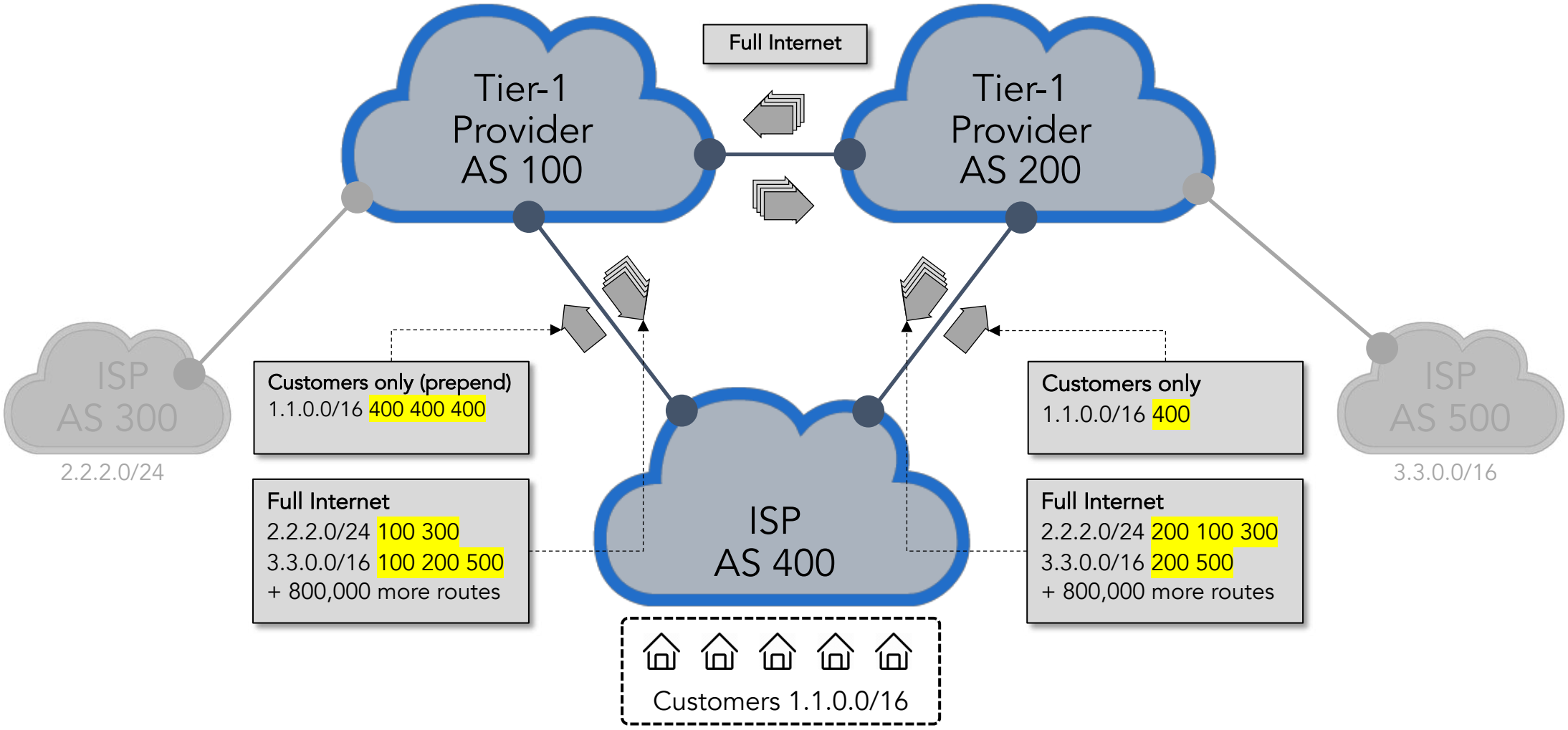
RIB-In, RIB-Local, RIB-Out, FIB



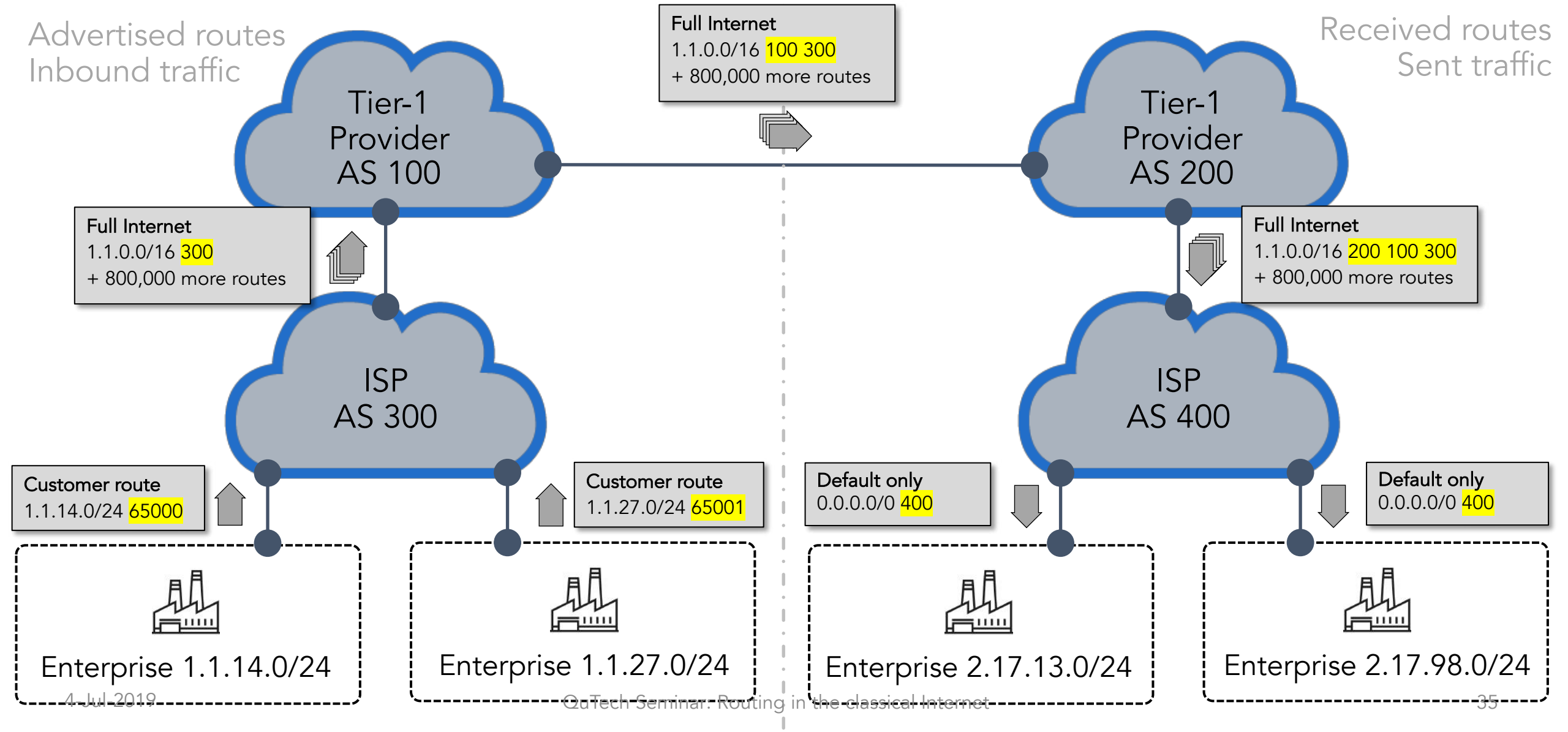
Example policy (Cisco)

```
router bgp 130
   network 121.10.0.0 mask 255.255.224.0
   neighbor 120.1.5.1 remote-as 120
   neighbor 120.1.5.1 prefix-list aggregate out
   neighbor 120.1.5.1 route-map routerD-out out
   neighbor 120.1.5.1 prefix-list default in
   neighbor 120.1.5.1 route-map routerD-in in
!
ip prefix-list aggregate permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
route-map routerD-out permit 10
  set as-path prepend 130 130 130
!
route-map routerD-in permit 10
  set local-preference 80
```

Example: Internet Service Provider (ISP)

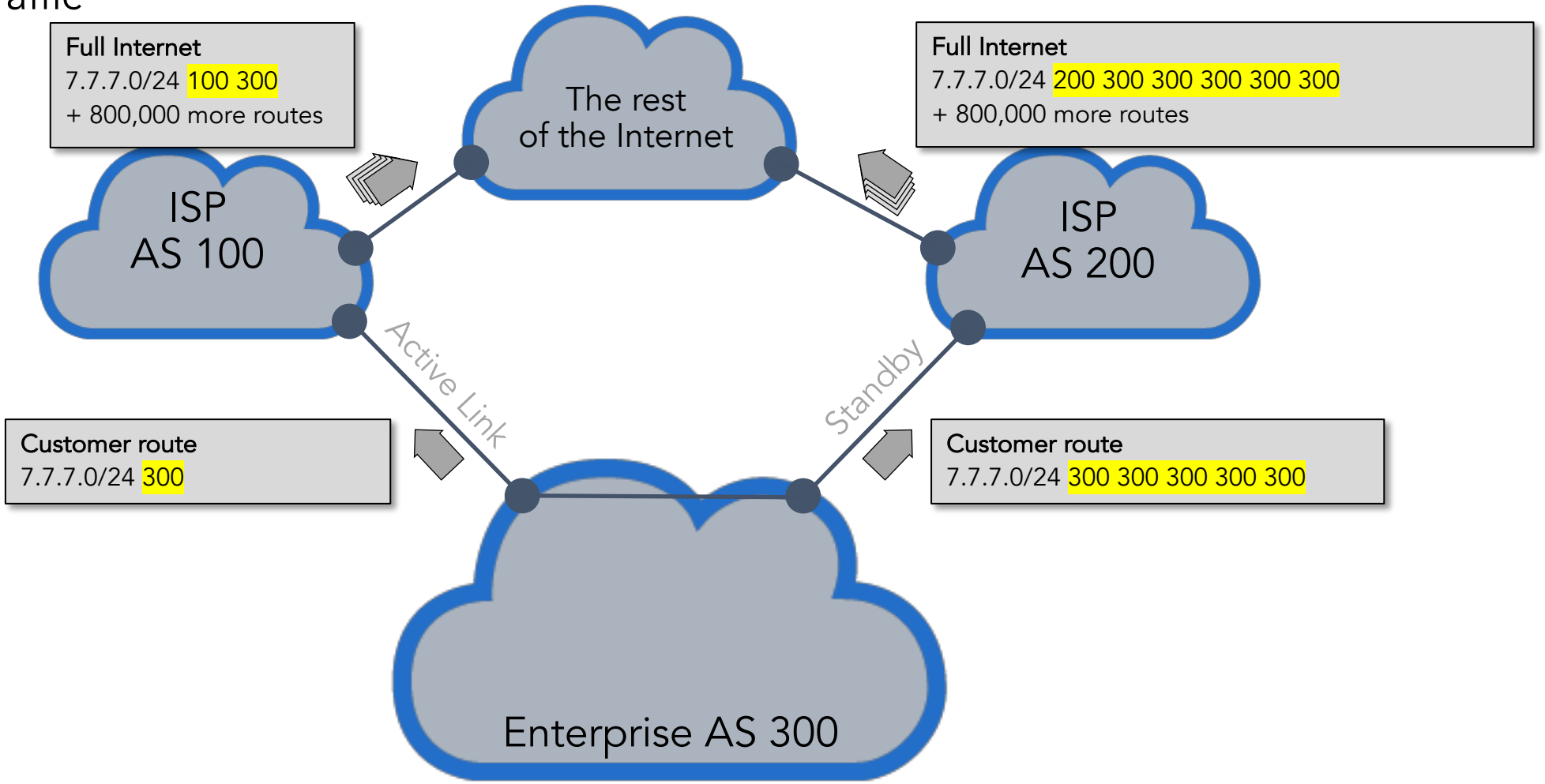


Single-homed enterprises



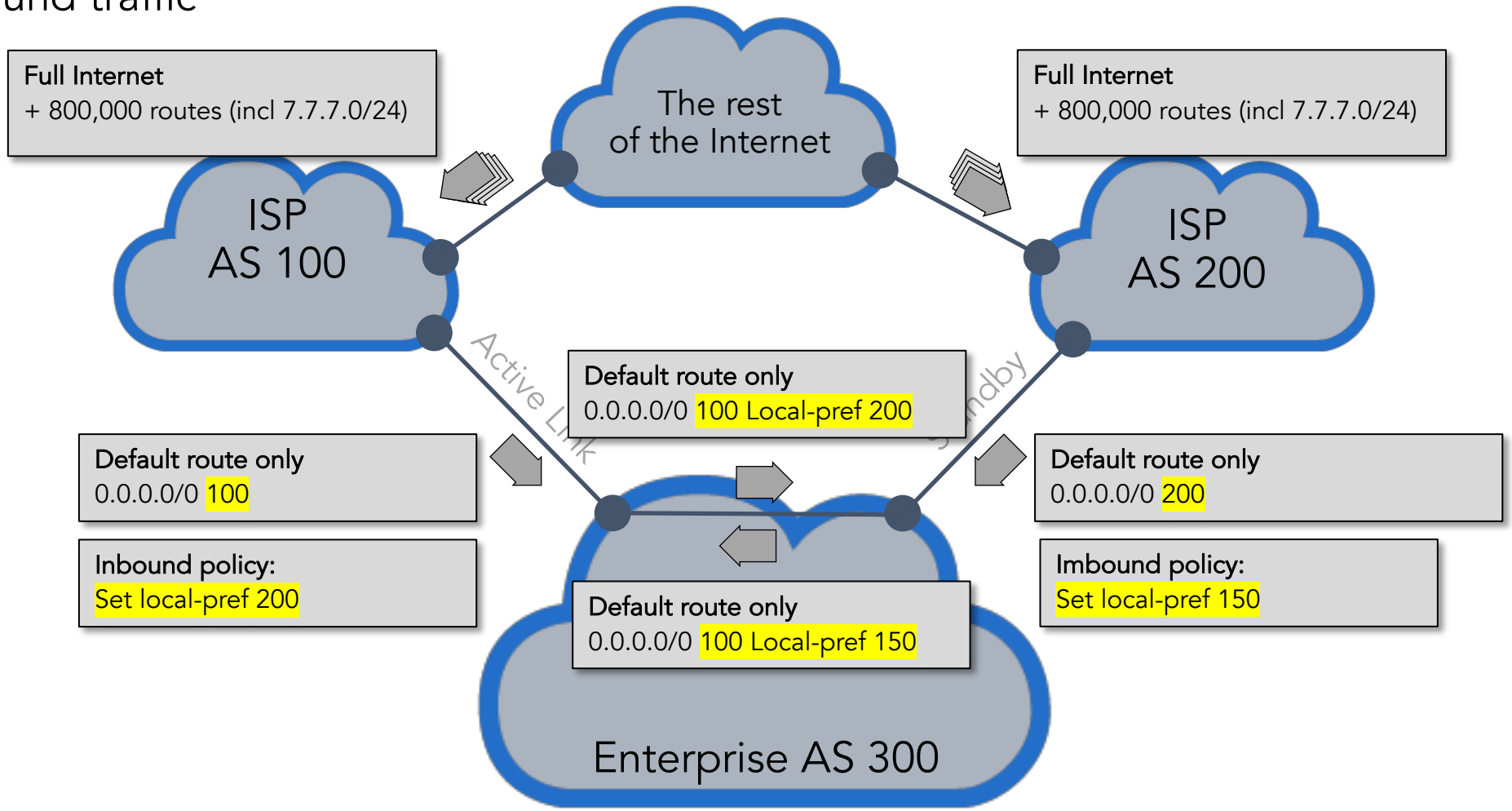
Dual-homed enterprise (active-standby)

Inbound traffic



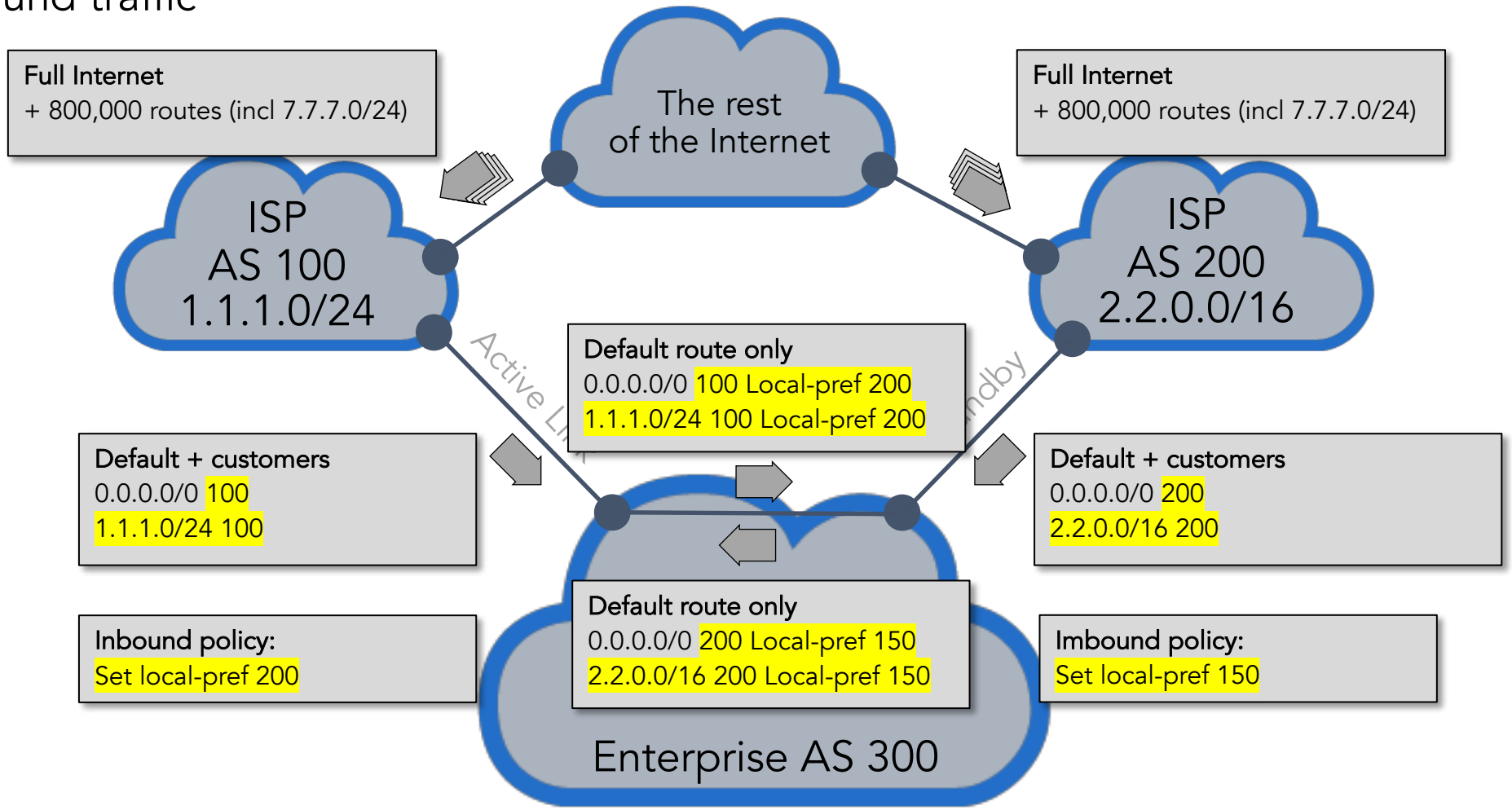
Dual-homed enterprise (active-standby)

Outbound traffic



Dual-homed enterprise (a-s optimized)

Outbound traffic



BGP security is severely lacking

The Register®
Biting the hand that feeds IT

Data Centre ▸ Networks

BGP super-blunder: How Verizon today sparked a 'cascading catastrophic failure' that knackered Cloudflare, Amazon, etc

'Normally you'd filter it out if some small provider said they own the internet'

By Kieren McCarthy in San Francisco 24 Jun 2019 at 19:01 61 SHARE ▼

Updated Verizon sent a big chunk of the internet down a black hole this morning – and caused outages at Cloudflare, Facebook, Amazon, and others – after it wrongly accepted a network misconfiguration from a small ISP in Pennsylvania, USA.

BANK INFO SECURITY®

Cryptocurrency Heist: BGP Leak Masks Ether Theft

Essential Internet Infrastructure - DNS, BGP - Remains Vulnerable, Experts Warn

computing

BGP route leak sends European mobile traffic via China

Yet another BGP hijack by China Telecom routes internet traffic of several European mobile operators via China

BGP trust model makes

Resource Public Key Infrastructure (RPKI) proposed to fix *some* of the problems.

Advanced BGP topics
(We just scratched the surface)

- BFD
- BGP in datacenters
- BGP over GRE
- BGP over MPLS
- Capability negotiation
- Communities
- Confederations
- Extended communities
- EVPN
- Flowspec
- Inter-AS VPNs a b c
- L2VPN
- L3VPN
- Labeled Unicast
- Link State BGP
- Link State BGP
- Multihoming
- Multipath
- Outbound Route Filters
- Redistribution
- Route reflectors
- Route servers
- RPKI
- SD-WAB
- Secure BGP
- Segment Routing
- YANG models
- Etc. etc. etc.…