Master Course
Computer Networks
IN2097

Prof. Dr.-Ing. Georg Carle

Chair for Network Architectures and Services
Department of Computer Science
Technische Universität München
http://www.net.in.tum.de

Outline

- Interdomain Routing
  - BGP: Border Gateway Protocol
  - Business considerations
    - Policy routing
    - Traffic engineering

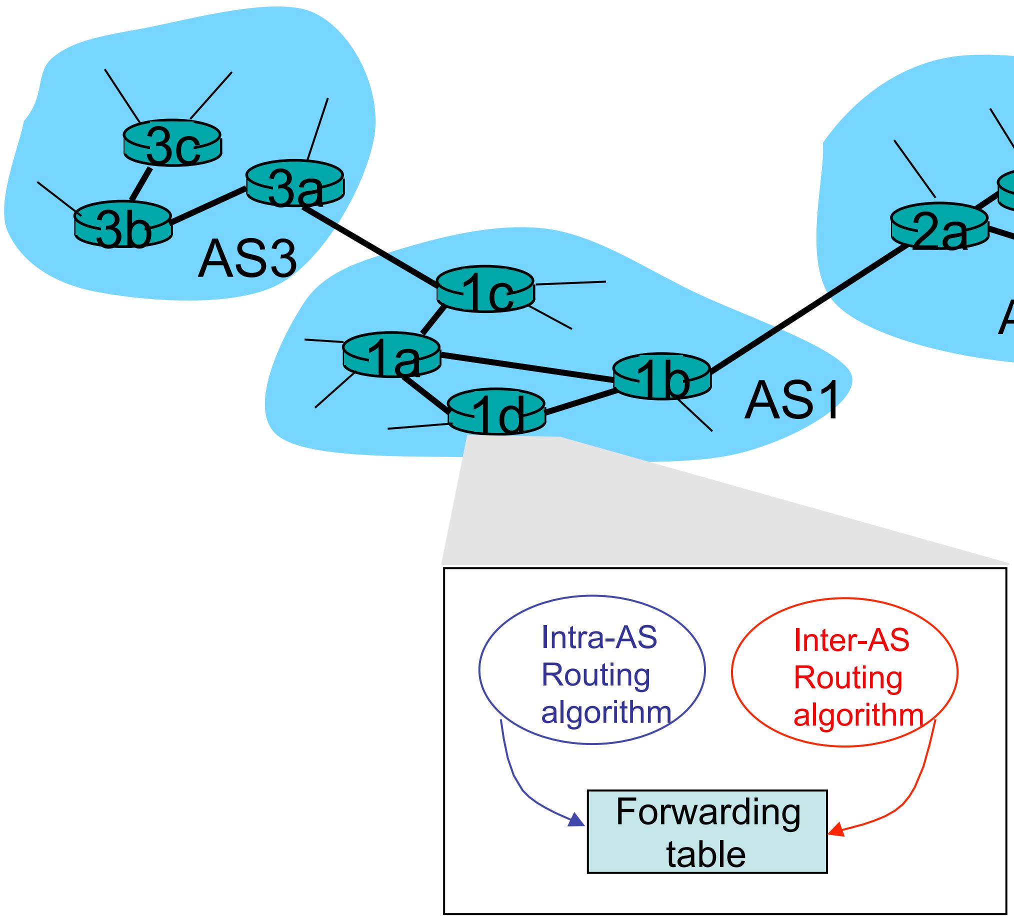
Interdomain Routing

BGP - Border Gateway Protocol

Hierarchical Routing

- Aggregate routers into regions called "autonomous systems" (short: AS; plural: ASes)
  - One AS ≈ one ISP / organisation
- Routers within one AS run same routing protocol
  - = "intra-AS" routing protocol (also called "intradomain")
  - Routers in different ASes can run different intra-AS routing protocols
- ASes are connected: via gateway routers
  - Direct link to [gateway] router in another AS
    = "inter-AS" routing protocol (also called "interdomain")
  - Warning: Non-gateway routers may need to know about inter-AS routing as well!

Interconnected ASes



- Forwarding table configured by both intra- *and* inter-AS routing algorithm:
  - Intra-AS sets entries for internal destinations
  - Inter-AS *and* intra-AS set entries for external destinations

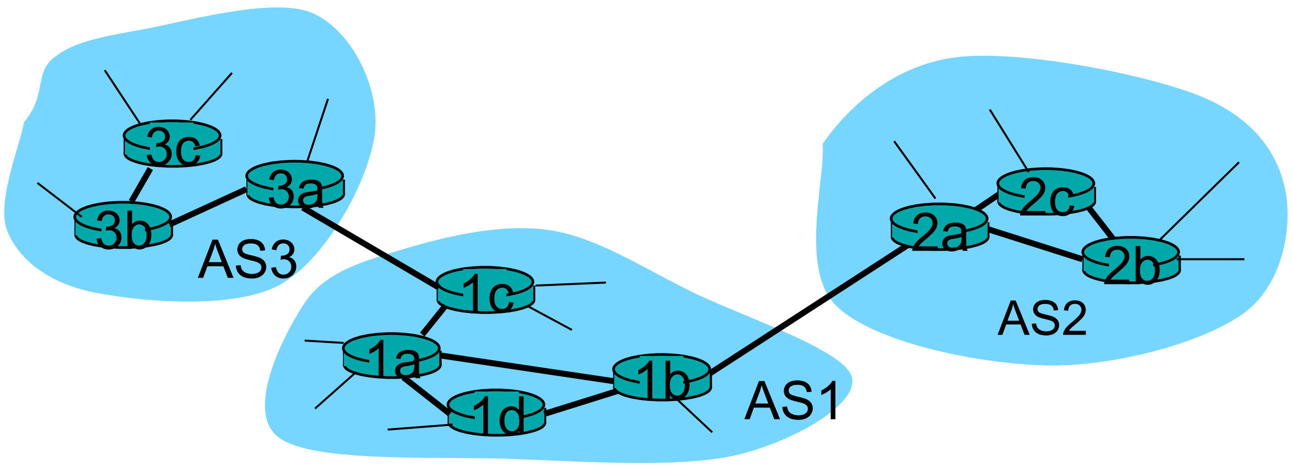
Inter-AS tasks

- Suppose router in AS1 receives datagram destined outside of AS1:
  - Router should forward packet to gateway router
  - …but to which one?

AS1 must:

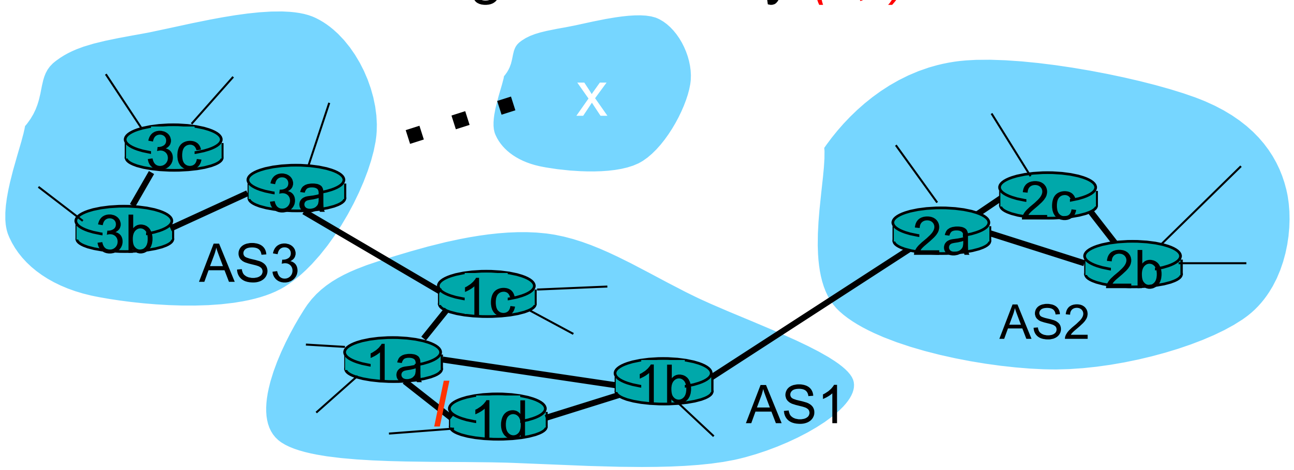
1. learn which destinations are reachable through AS2, which through AS3
2. propagate this reachability info *to all* routers in AS1 (i.e., not just the gateway routers)

Job of inter-AS routing!



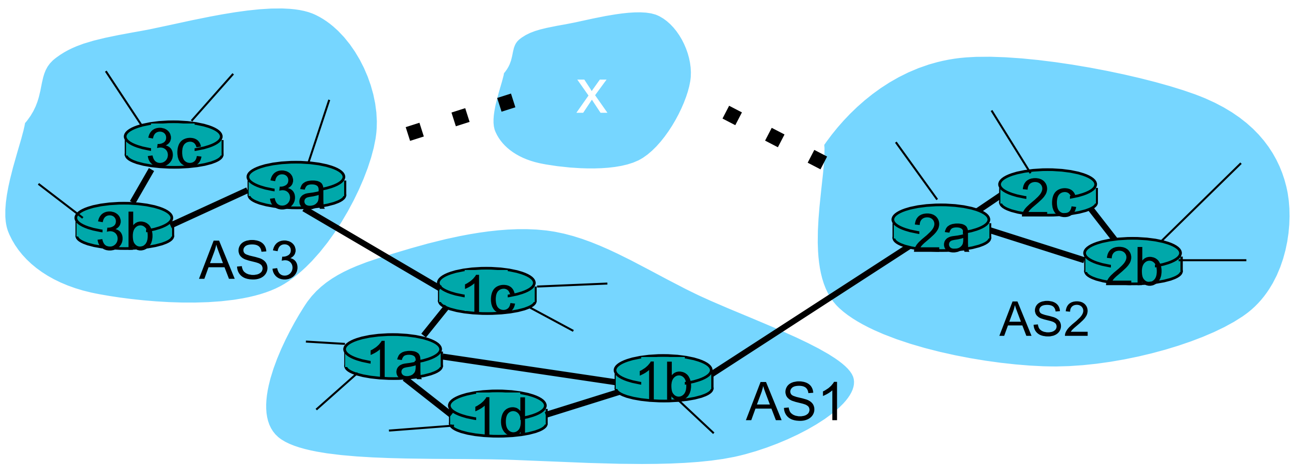
Example: Setting Forwarding Table in Router 1d

- Suppose AS1 learns (via inter-AS protocol) that subnet *x* is reachable via AS3 (gateway 1c) but not via AS2.
- Inter-AS protocol propagates reachability info to all internal routers.
- Router 1d determines from intra-AS routing info that its interface *I* (i.e., interface to 1a) is on the least cost path to 1c.
  - installs forwarding table entry *(x,I)*



Example: Choosing among multiple ASes

- Now suppose AS1 learns from inter-AS protocol that subnet x is reachable from AS3 and from AS2.
- To configure forwarding table, router 1d must determine towards which gateway it should forward packets for destination x.
  - "Do we like AS2 or AS3 better?"
  - Also the job of inter-AS routing protocol!



Interplay of Inter-AS and Intra-AS Routing

- Inter-AS routing
  - Only for destinations outside of own AS
  - Used to determine gateway router
  - Also: Steers transit traffic
    (from AS *x* to AS *y* via our own AS)
- Intra-AS routing
  - Used for destinations within own AS
  - Used to reach gateway router for destinations outside own AS

⇒ Often, routers need to run *both* types of routing protocols… even if they are not directly connected to other ASes!

Internet inter-AS routing: BGP

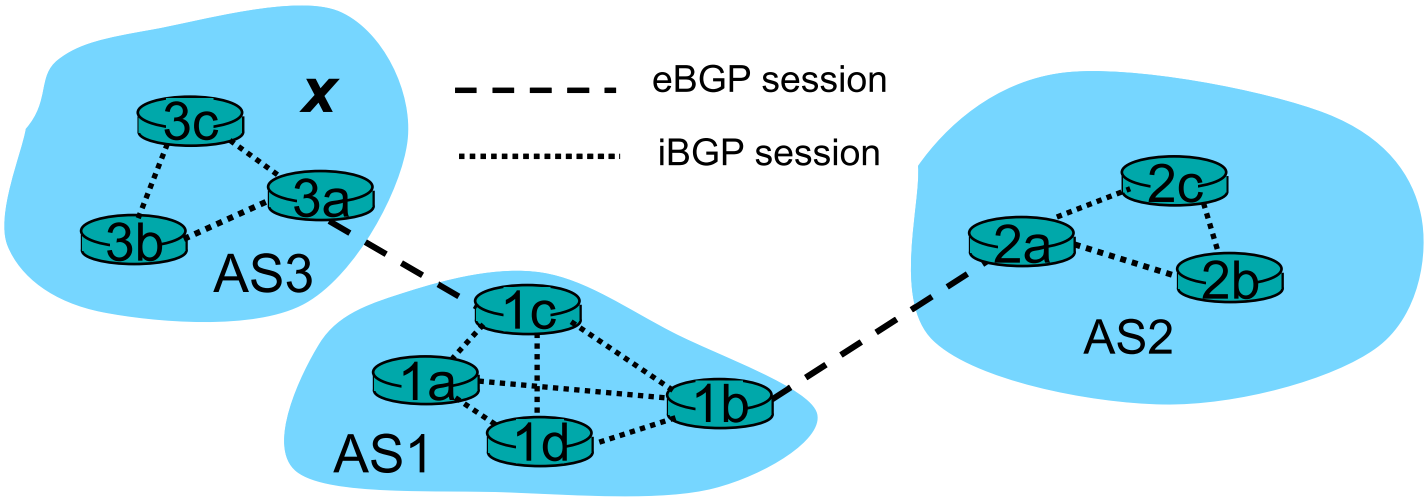
- BGP (Border Gateway Protocol):
  The de facto standard for inter-AS routing
- BGP provides each AS a means to:
  1. Obtain subnet reachability information from neighboring ASes
  2. Propagate reachability information to all AS-internal routers
  3. Determine "good" routes to subnets based on reachability information and policy
- Allows an AS to advertise the existence of an IP prefix to rest of Internet: *"This subnet is here"*

BGP Basics

- Pairs of routers (BGP peers) exchange routing info over semi-permanent TCP connections:
  BGP sessions
  - BGP sessions need not correspond to physical links!
- When AS2 advertises an IP prefix to AS1:
  - AS2 *promises* it will forward IP packets towards that prefix
  - AS2 can aggregate prefixes in its advertisement
    (e.g.: 10.11.12.0/26, 10.11.12.64/26, 10.11.12.128/25
    into 10.11.12.0/24)

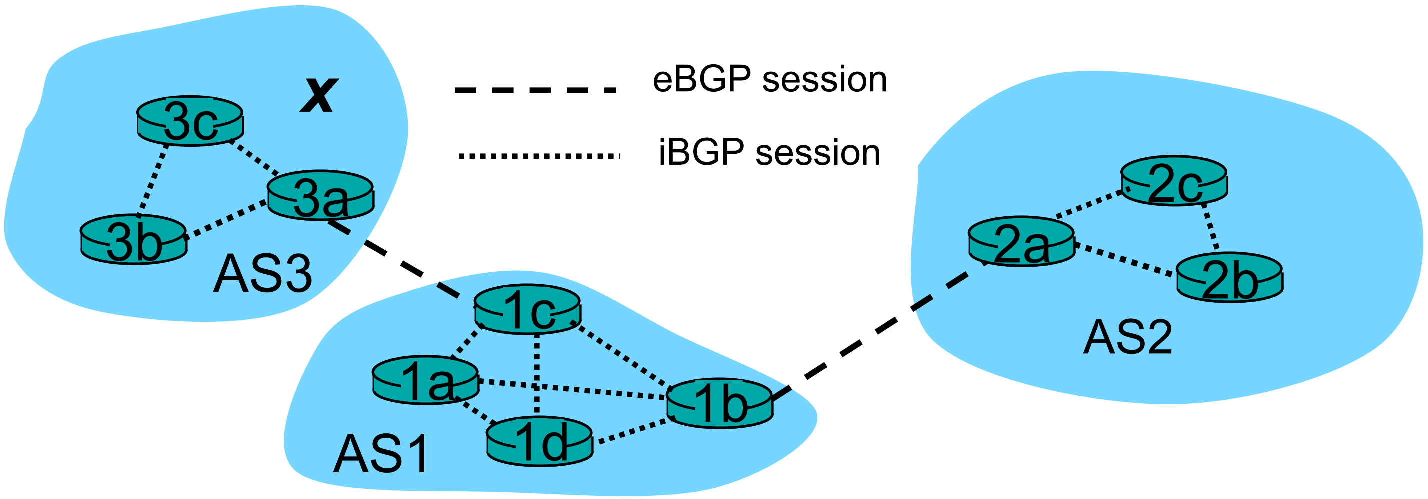
eBGP and iBGP

- External BGP: between routers in *different* ASes
- Internal BGP: between routers in *same* AS
  - Remember: In spite of intra-AS routing protocol, *all* routers need to know about external destinations (not only border routers)
- Not different protocols—just slightly different configurations!



Distributing reachability info

- Using eBGP session between 3a and 1c, AS3 sends reachability information about prefix *x* to AS1.
  - 1c can then use iBGP to distribute new prefix information to all routers in AS1
  - 1b can then re-advertise new reachability information to AS2 over 1b-to-2a eBGP session
- When router learns of new prefix *x*, it creates entry for prefix in its forwarding table.



AS Numbers

- How do we express a BGP path?
- ASes identified by *AS Numbers* (short: ASN)
  Examples:
  - Leibnitz-Rechenzentrum = AS12816
  - Deutsche Telekom = AS3320
  - AT&T = AS7018, AS7132, AS2685, AS2686, AS2687
- ASNs used to be 16bit, but also have 32bit nowadays
  - May have problems with 16bit ASNs on very old routers
- ASN assignment: similar to IP address space
  - ASN space administered IANA
  - Local registrars, e.g., RIPE NCC in Europe

Path attributes & BGP routes

- Advertised prefix includes [many] BGP attributes
  - prefix + attributes = “route”
- Most important attributes:
  - AS-PATH: contains ASes through which prefix advertisement has passed: e.g., AS 67, AS 17, AS 7018
  - NEXT-HOP: indicates specific internal-AS router to next-hop AS (may be multiple links from current AS to next-hop-AS)
- When gateway router receives route advertisement, it uses an import policy to accept/decline the route
  - More on this later

How does BGP work?

- BGP = “path++” vector protocol
- BGP messages exchanged using TCP
  - Possible to run eBGP sessions not on border routers
- BGP message types:
  - OPEN: set up new BGP session, after TCP handshake
  - NOTIFICATION: an error occurred in previous message → tear down BGP session, close TCP connection
  - KEEPALIVE: “null” data to prevent TCP timeout/auto-close; also used to acknowledge OPEN message
  - UPDATE:
    - Announcement: inform peer about new / changed route to some target
    - Withdrawal: (inform peer about non-reachability of a target)

BGP updates

- Update (Announcement) message consists of
  - Destination (IP prefix)
  - AS Path (=Path vector)
  - Next hop (=IP address of our router connecting to other AS)
- …but update messages also contain a lot of further attributes:
  - Local Preference: used to prefer one gateway over another
  - Origin: route learned via { intra-AS | inter-AS | unknown }
  - MED (MULTI_EXIT_DISC): used on external (inter-AS) links to discriminate among multiple exit or entry points
  - Community: tags applied to prefixes for common treatment

⇒ Not a pure path vector protocol: More than just the path vector

- Local configuration uses much more information than what is exchanged in messages

⇒ BGP is an "information hiding protocol" (quote from Randy Bush)

BGP update: Very simple example

- Type: Announcement
  - Either this is a new route to the indicated destination,
  - or the existing route has been changed
- Destination prefix: 10.11.128.0/17
- AS Path:

`7018 3320 4711 815 12816`

Current AS (7018)

Originator: The AS that "owns" 10.11.128.0/17 (12816)

- Next Hop: 192.168.69.96
  - The router that connects the current AS to AS 3320

How the update travelled ←

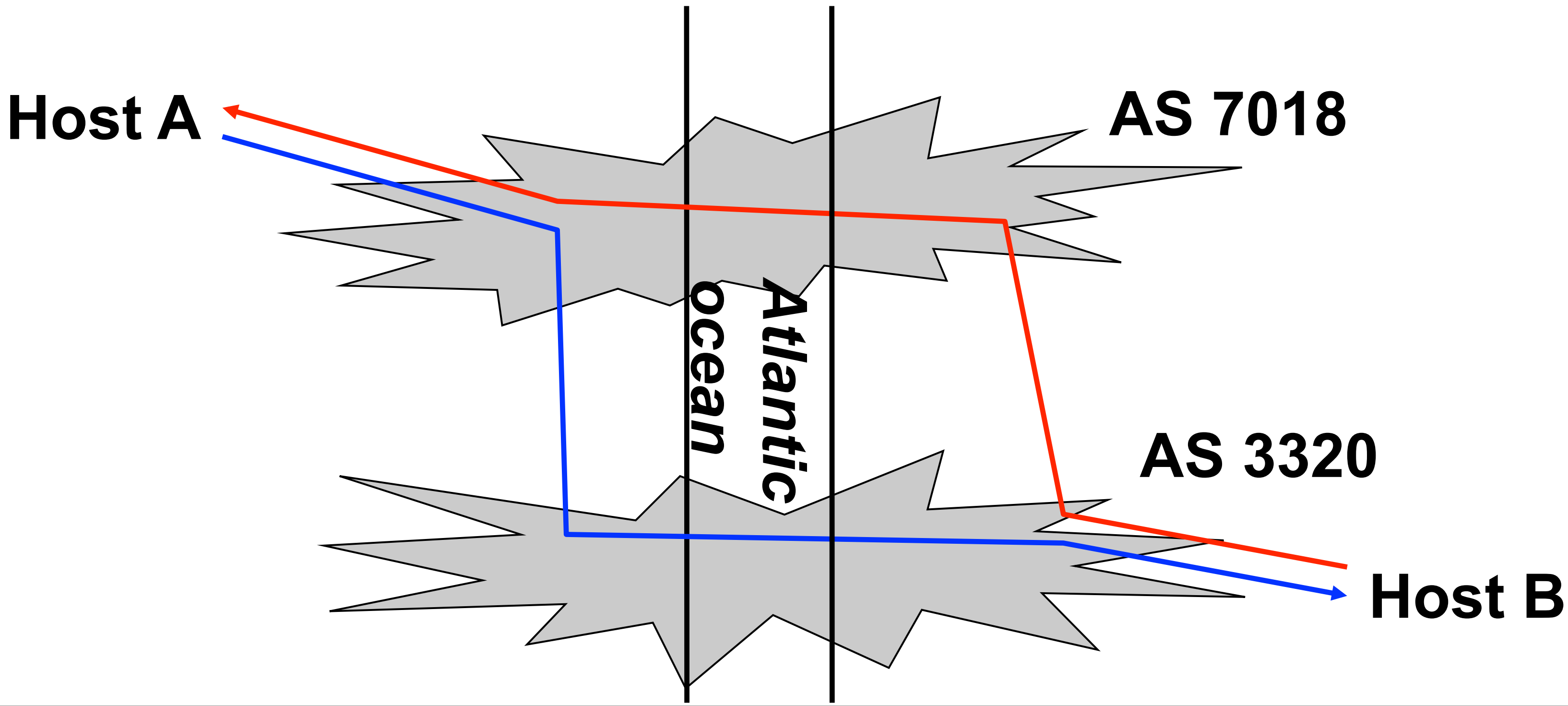
→ **How the IP packets will be forwarded** (if this route gets chosen)

BGP route selection

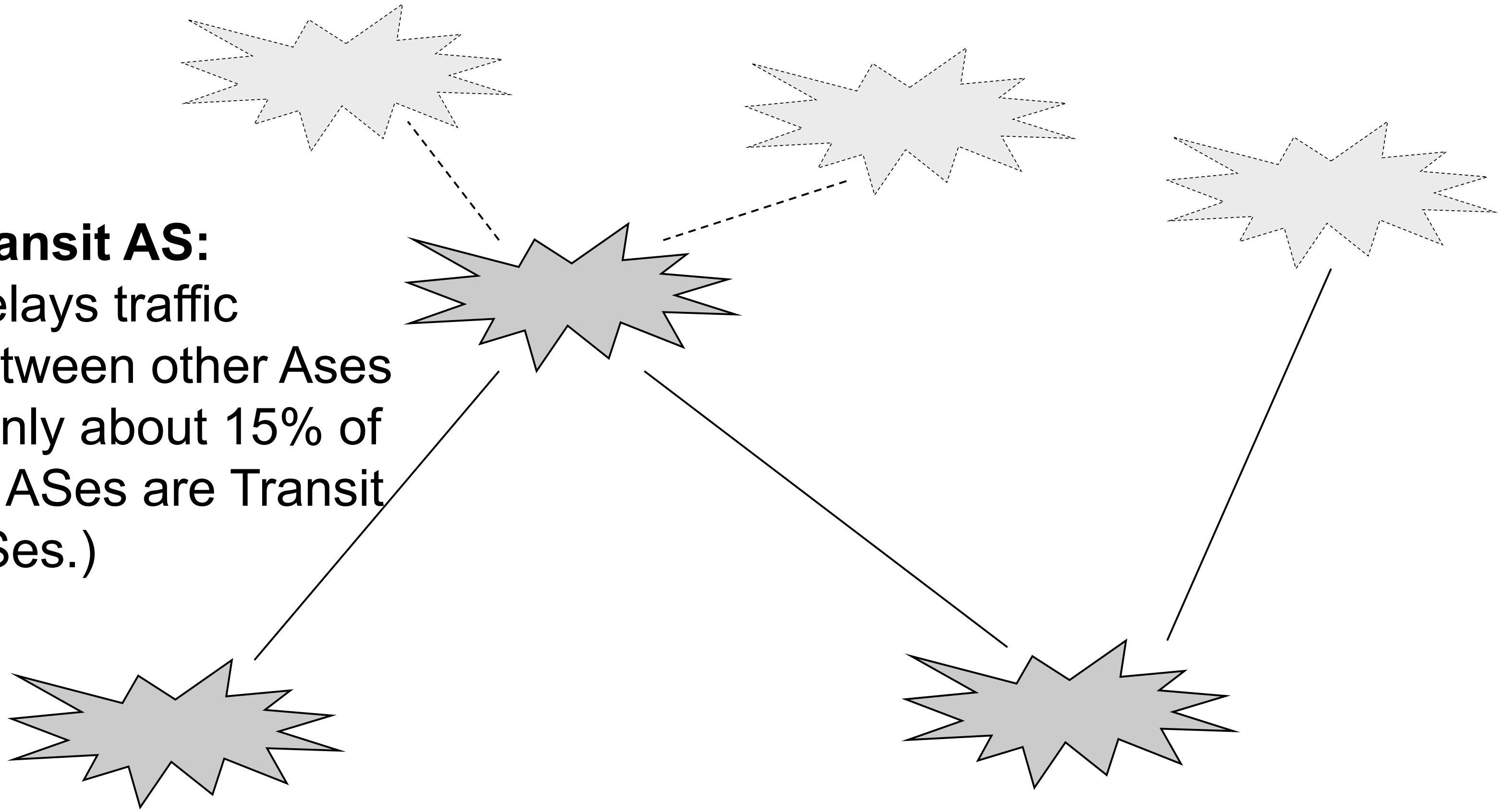
- Router may learn about more than 1 route to some prefix
  ⇒ Router must select the best one among these
- Elimination rules (**simplified**):
  1. Local preference value attribute: policy decision
  2. Shortest AS-PATH
  3. Closest NEXT-HOP router outside AS: hot potato routing
  4. Additional criteria

Business and Hot-potato routing

- Interaction between Inter-AS and Intra-AS routing
  - Business: If traffic is destined for other AS, get rid of it ASAP
  - Technical: Intra-AS routing finds shortest path to gateway
- Multiple transit points ⇒ asymmetrical routing
  - Asymmetrical paths are very common on the Internet



Terminology: Transit AS, Stub AS, Multi-homed AS



Transit AS:
Relays traffic between other Ases (Only about 15% of all ASes are Transit ASes.)

Stub AS: Buys transit from only one other AS, but does not offer transit for other ASes

Multi-homed AS: Buys transit from ≥2 other ASes, but does not offer transit for other ASes

Business relationships

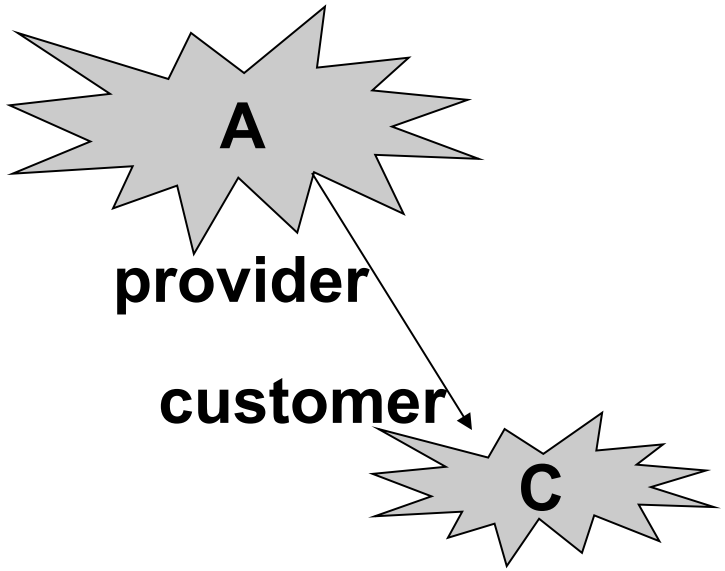
- Internet = network of networks (ASes)
  - Many thousands of ASes
  - Not every network connected to every other network
  - BGP used for routing between ASes
- Differences in economical power/importance
  - Some ASes huge, intercontinental (AT&T, Cable&Wireless)
  - Some ASes small, local (e.g., München: M-Net, SpaceNet)
- Small ASes customers of larger ASes: Transit traffic
  - Smaller AS pays for connecting link + for data = buys transit
  - Business relationship = customer—provider
- Equal-size/-importance ASes
  - Usually share cost for connecting link[s]
  - Business relationship = peering (*specific* transit traffic is for free)
- Warning: peering ("equal-size" AS)
  ≠ peers of a BGP connection (also may be customer or provider)
  ≠ peer-to-peer network

Business and policy routing (1)

- Basic principle #1 (Routing)
  - Prefer routes that incur financial gain
- Corollary: If you have the choice, then…
  - …routes via a customer…
  - …are better than routes via a peer, which…
  - …are better than routes via a provider.
- Basic principle #2 (Route announcement)
  - Announce routes that incur financial gain if others use them
    - Others = customers
  - Announce routes that reduce costs if others use them
    - Others = peers
  - Do not announce routes that incur financial loss (…as long as alternative paths exist)

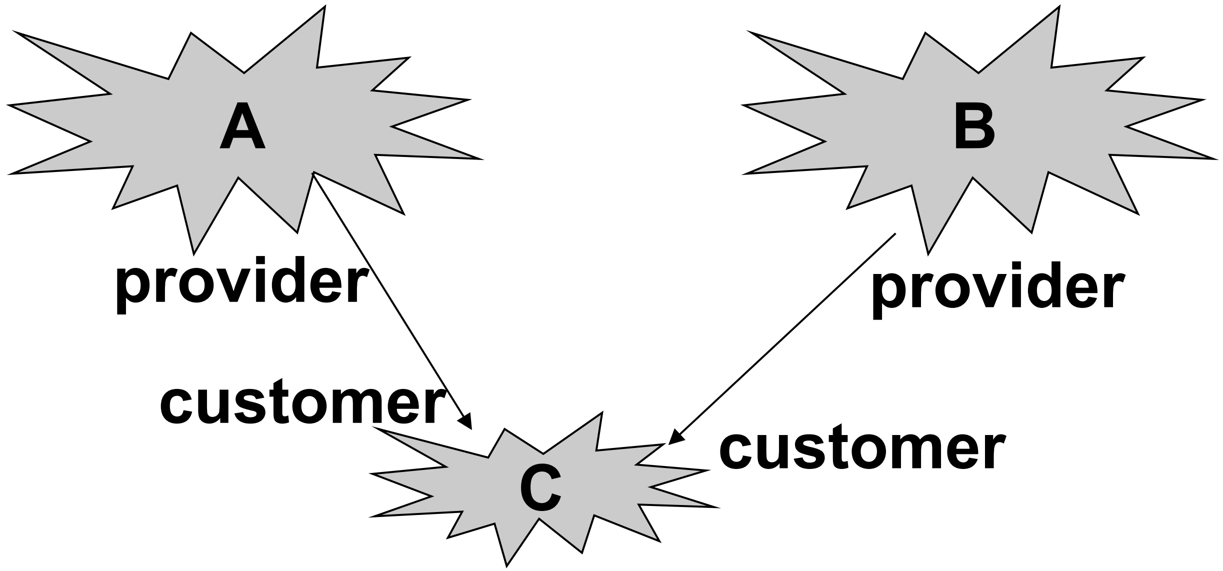
Business and policy routing (2)

- A tells C all routes it uses to reach other ASes
  - The more traffic comes from C, the more money A makes



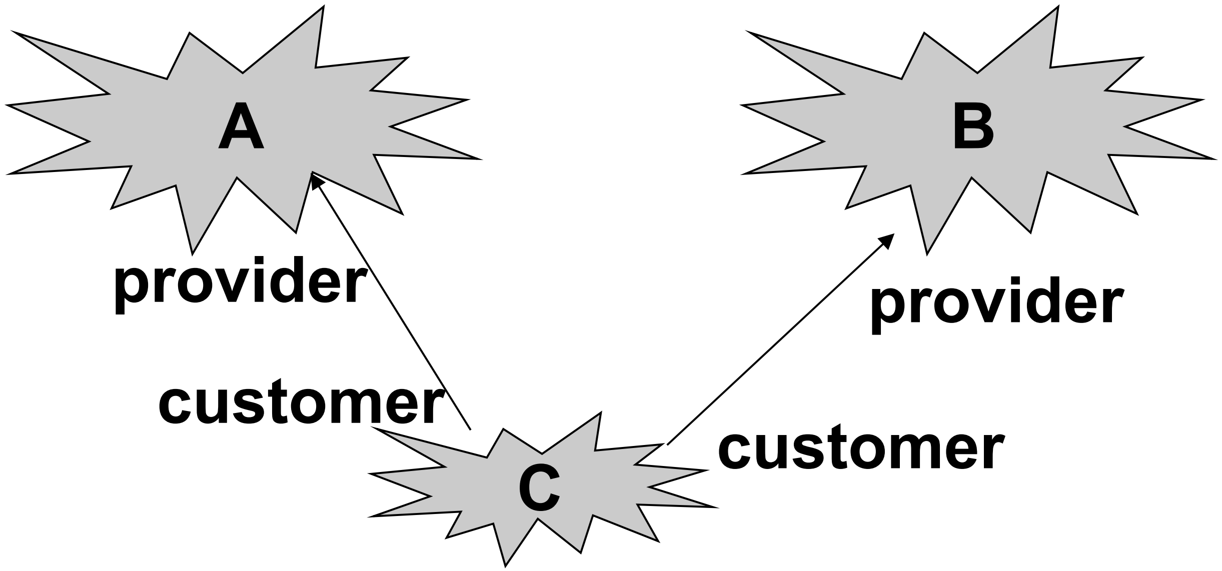
Business and policy routing (3)

- A and B tell C all routes they use to reach other ASes
  - The more traffic flows from C to A, the more money A makes
  - The more traffic flows from C to B, the more money B makes
  - C will pick the one with the cheaper offer / better quality / …



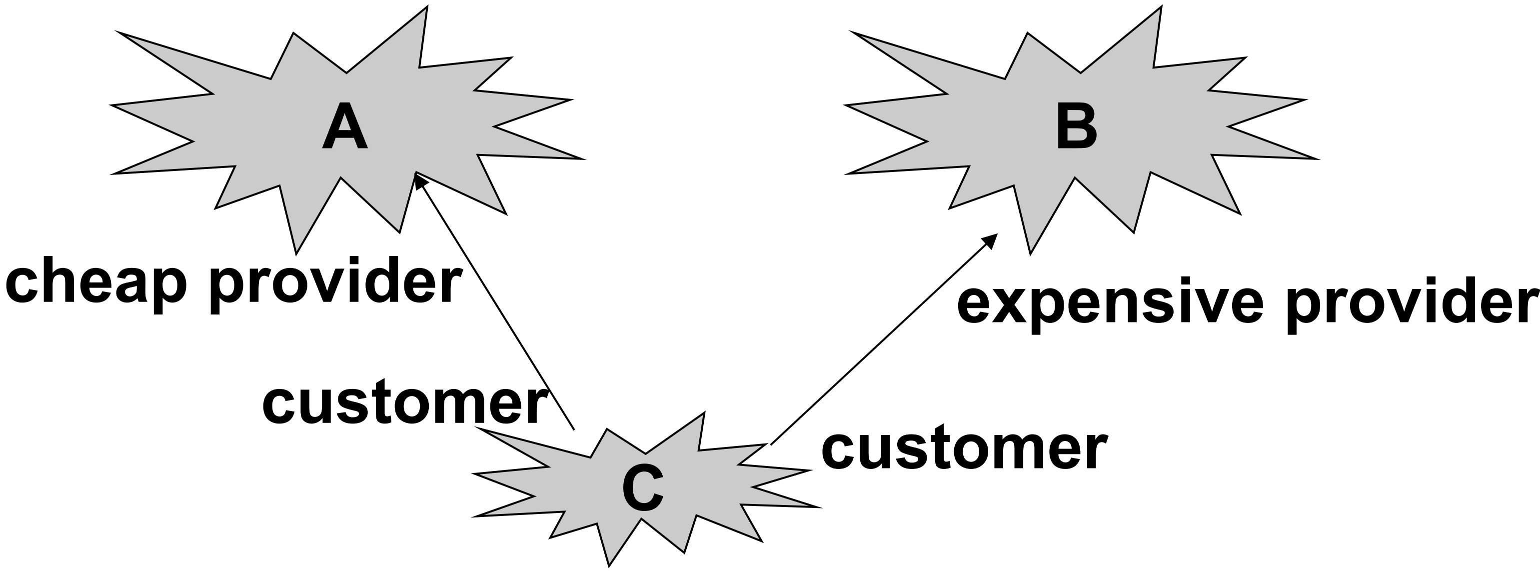
Business and policy routing (4)

- C tells A its own prefixes; C tells B its own prefixes
  - C wants to be reachable from outside
- C does not tell A routes learned from/via B
  C does not tell B routes learned from/via A
  - C does not want to pay money for traffic …↔A ↔C ↔B ↔…



Business and policy routing (5): AS path prepending

- C tells A its own prefixes
- C may tell B its own prefixes
  - …but inserts "C" multiple times into AS path. Why?
  - Result: Route available, but longer path = less attractive
  - Technique is called *AS path prepending*

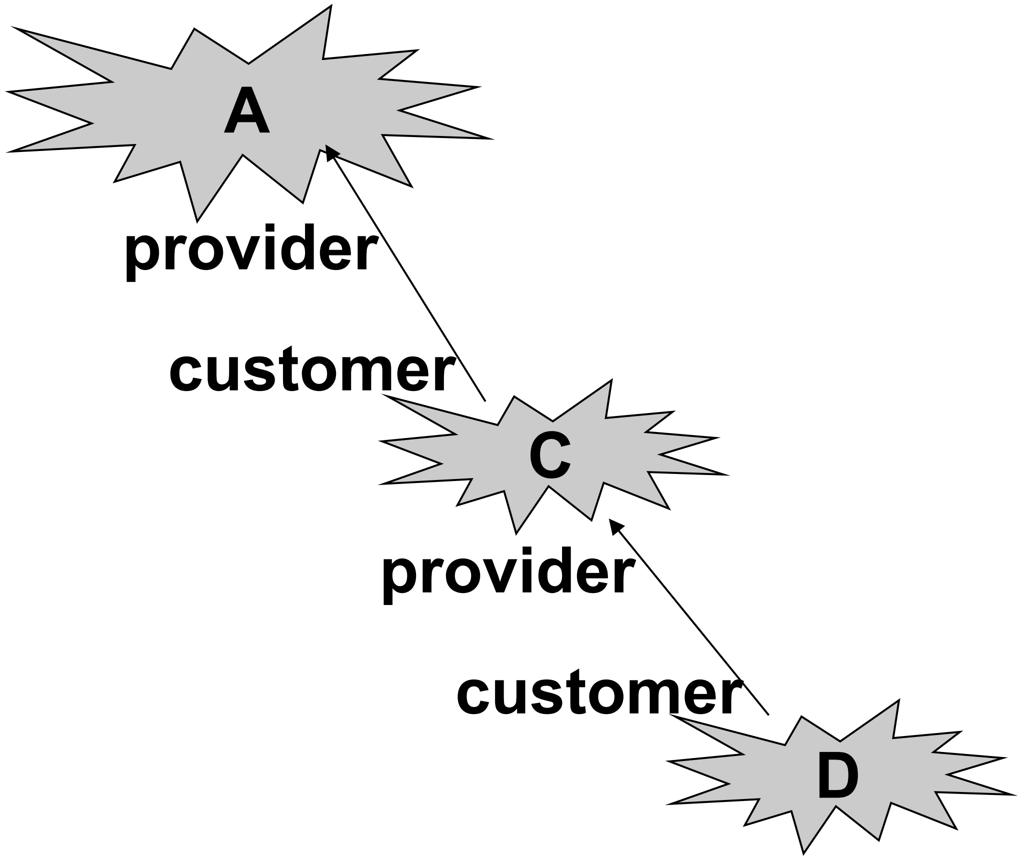


AS path prepending

- The same ASN *subsequently* within an AS path does not constitute a loop
- Recall the elimination rule for selecting from multiple path alternatives
  - "Prefer the shortest AS path" is rule 2
  - Only ignored if *Local Pref* value is set
  - AS path prepending makes a route less attractive – will then only be used when there is no alternative
- How many times to repeat the AS number?
  - Usually just 1 or 2 repetitions
  - More than ≈5 is useless

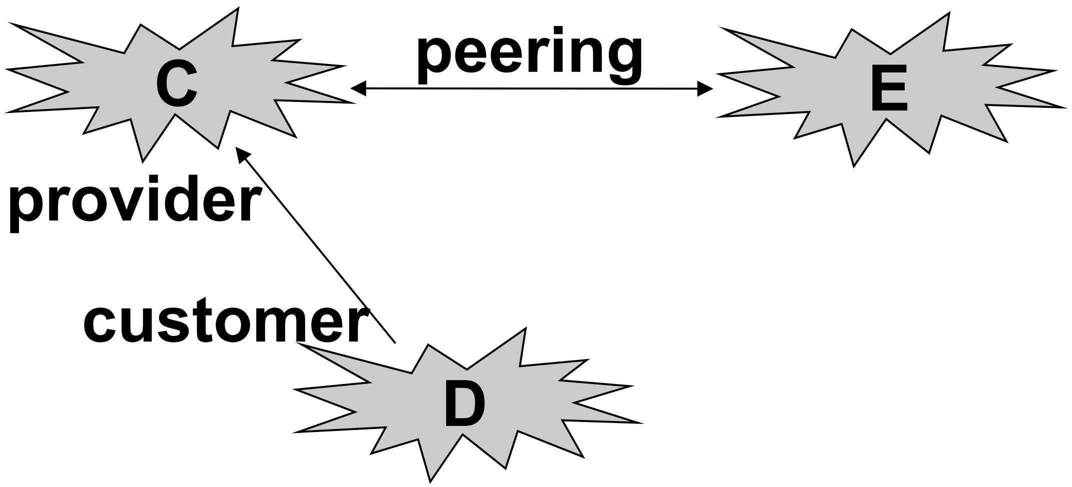
Business and policy routing (6)

- What should C announce here?
  - C tells A about its own prefixes
  - C tells A about its route to D's prefixes:
    loses money to A, but gains money from D



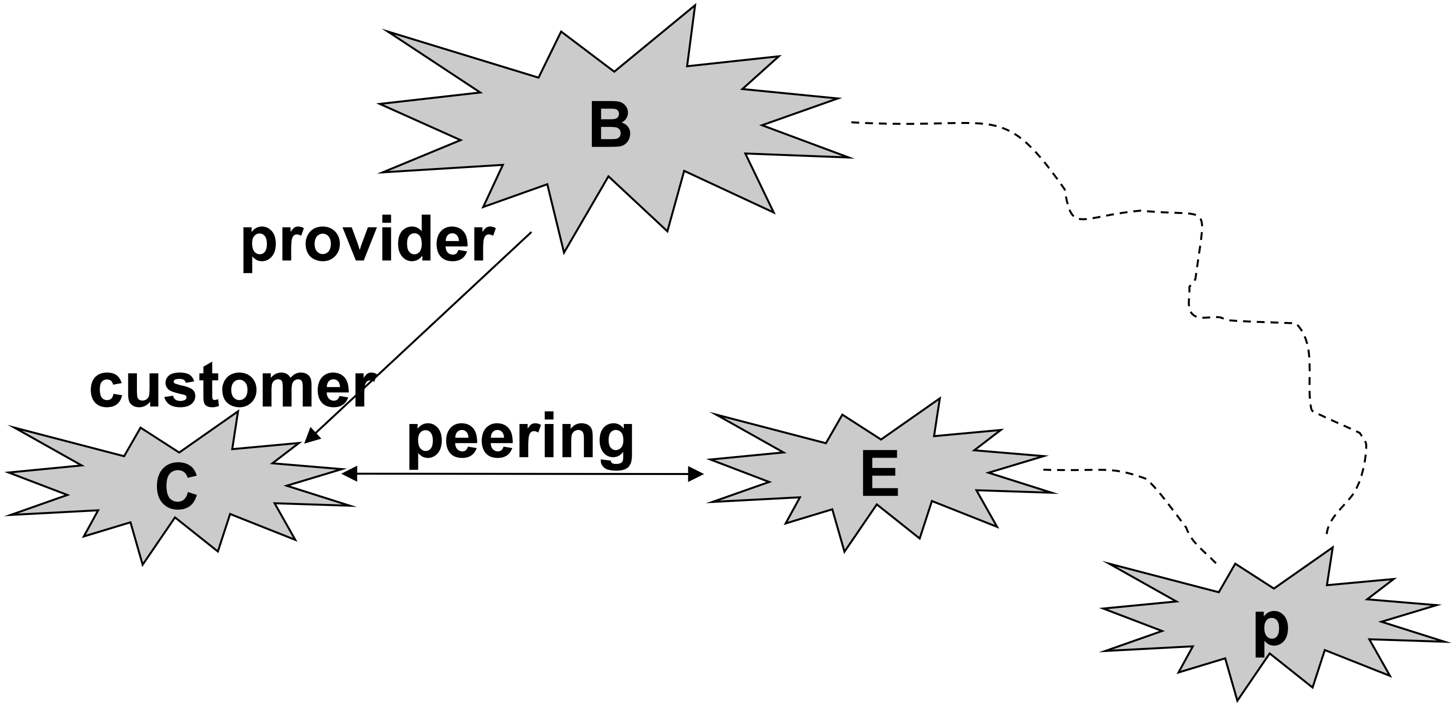
Business and policy routing (7)

- What should C announce here?
  - C tells peering partner E about its own prefixes and route to D:
    no cost on link to E, but gains money from D



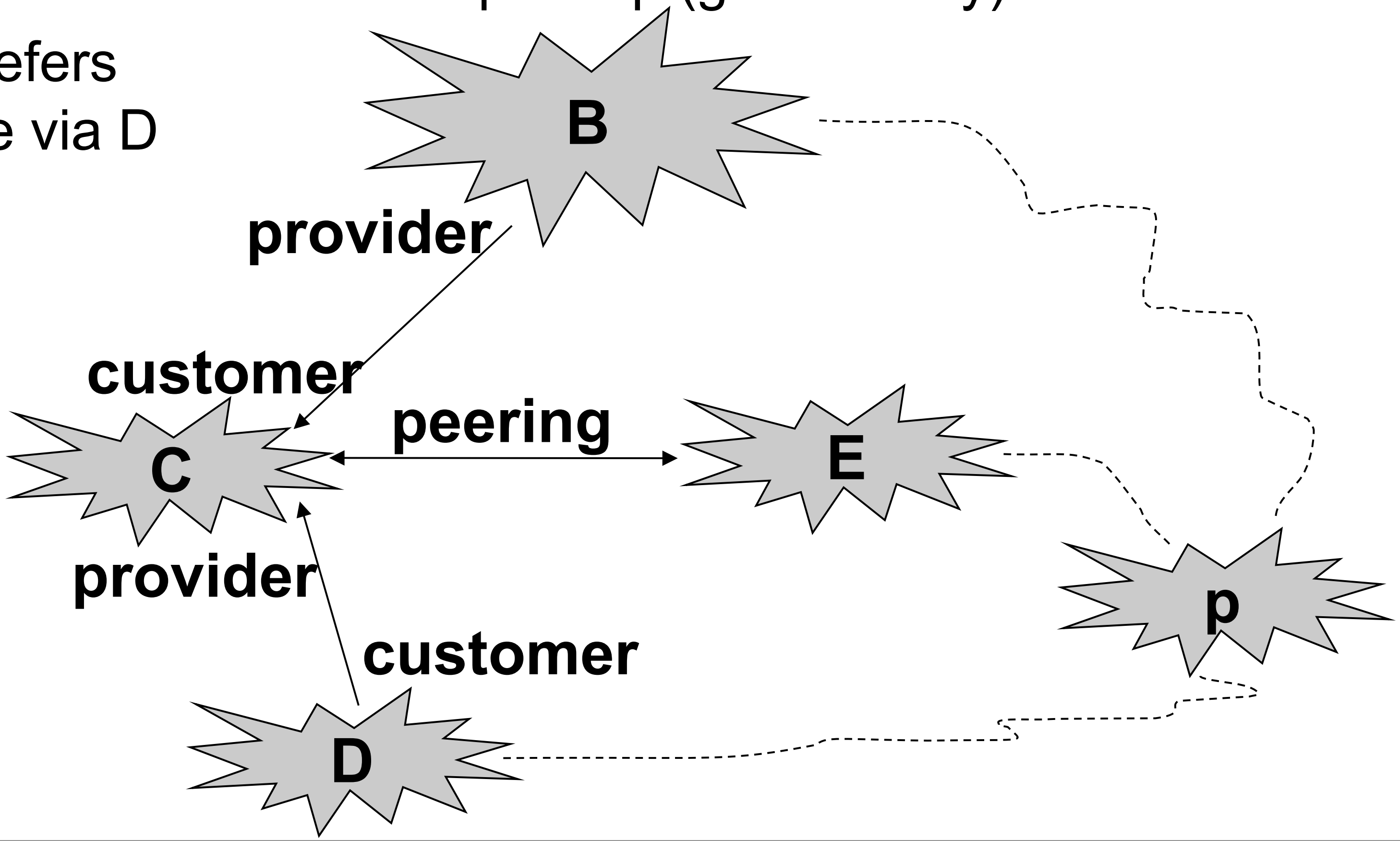
Business and policy routing (8a)

- Which route should C select?
  - B tells C about route to prefix p (lose money)
  - E tells C about route to prefix p (± 0)
  - C prefers route via E



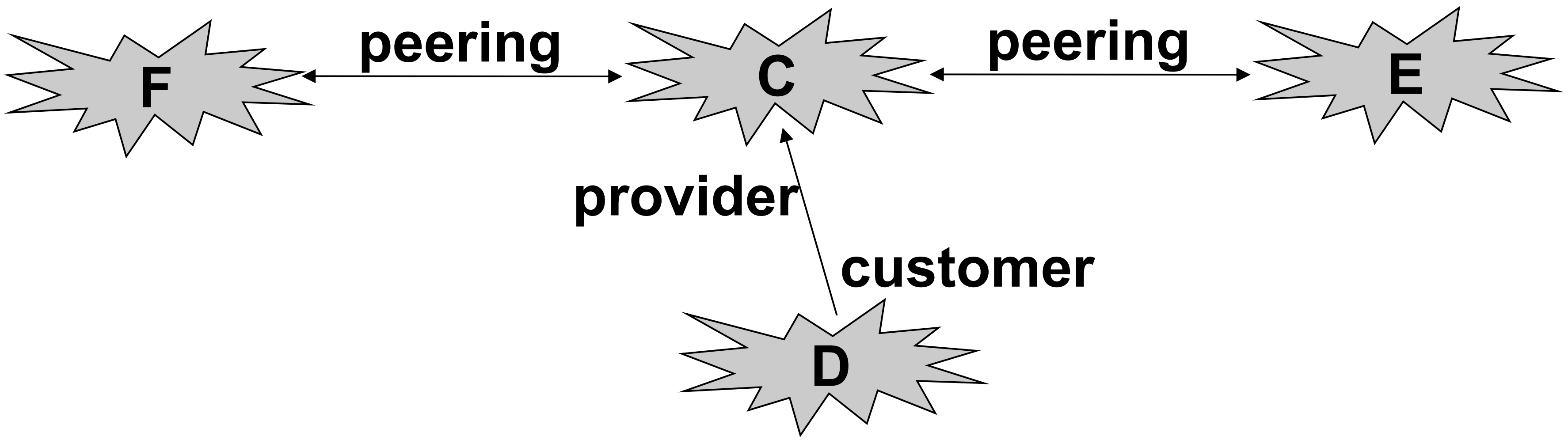
Business and policy routing (8b)

- Which route should C select?
  - B tells C about route to prefix p (lose money)
  - E tells C about route to prefix p (± 0)
  - D tells C about route to prefix p (gain money)
  - C prefers route via D



Business and policy routing (9)

- What should C announce here?
  - C announces to F and E: its own prefixes and D’s routes
  - C does *not* announce to E: routes going via F
    - Otherwise: E could send traffic towards F but wouldn’t pay anything, F wouldn’t pay either, and C’s network gets loaded with additional traffic
  - C does *not* announce to F: routes going via E
    - Same reason

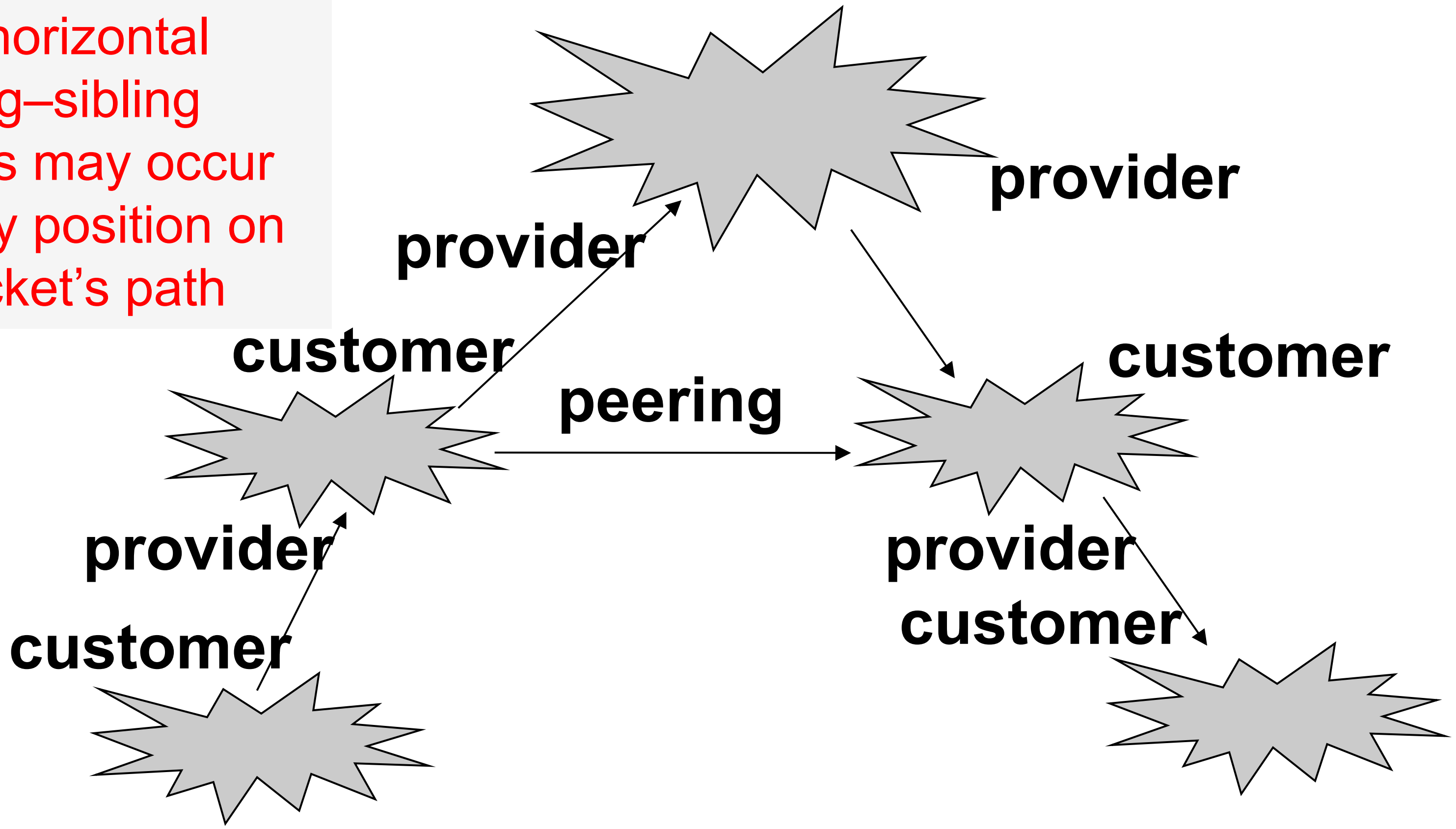


Policy routing: Valley-Free Routing (Idealised)

Results: Packets always travel…

1. upstream: sequence of C→P links (possibly length = 0)
2. then possibly across *one* peering link
3. then downstream: sequence of P→C links (possibly length = 0)

But: horizontal sibling–sibling edges may occur at any position on a packet's path

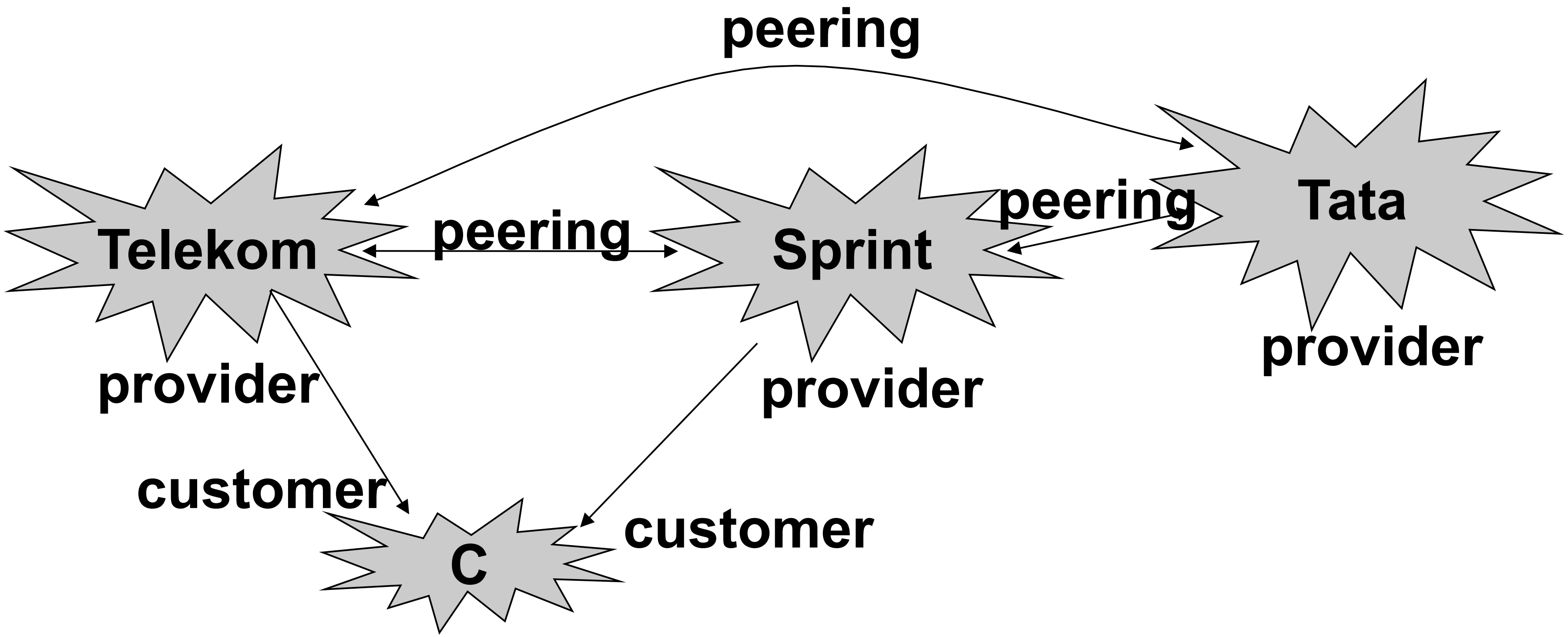


Siblings

- Not everything is provider/customer or peering
- Sibling = mutual transit agreement
  - Provide connectivity to the rest of the Internet for each other
  - ≈ very extensive peering
- Examples
  - Two small ASes close to each other that cannot / do not want to afford additional Internet services
  - Merging two companies
    - Merging two ASes into one = difficult,
    - Keeping two ASes and exchanging everything for free = easier
  - Example: AT&T has five different AS numbers (7018, 7132, 2685, 2686, 2687)

Business and policy routing (10): “Tiers” / “DFZ”

- Big players have no providers, only customers and peers
  - “Tier-1” ISPs
  - or “Default-Free Zone” (DFZ - have no default route to a “provider”)
- Each Tier-1 peers with each other



Tier-1, Tier-2, Tier-3 etc.

- Tier-1/DFZ = only peerings, no providers
- Tier-2 = only peerings and one or more Tier-1 providers
- Tier-3 = at least one Tier-2 as a provider
- Tier-*n* = at least one Tier-(n-1) provider
  - defined recursively
  - *n*≥4: Rare in Western Europe, North America, East Asia

- “Tier-1.5” = almost a Tier-1 but pays money for *some* links
  - Example: Deutsche Telekom used to pay money to Sprint, but is now Tier-1
  - Marketing purposes: Tier-1 sounds better

BGP Policy Routing: Technical summary

1. Receive BGP update
2. Apply import policies
   - Filter routes
   - Tweak attributes (advanced topic …)
3. Best route selection based on attribute values
   - Policy: Local Pref settings and other attributes
   - Install forwarding tables entries for best routes
   - Possibly transfer to Route Reflector
     (RR is alternative to logical full mesh of iBGP sessions)
4. Apply export policies
   - Filter routes
   - Tweak attributes
5. Transmit BGP updates

BGP policy routing: Business relationship summary

- Import Policy = Which routes to use
  - Select path that incurs most money
  - Special/political considerations (e.g., Iranian AS does not want traffic to cross Israeli AS; other kinds of censorship)
- Export Policy = Which routes to propagate to other ASes
  - Not all known routes are advertised:
    Export only…
    - If it incurs revenue
    - If it reduces cost
    - If it is inevitable
- **Policy routing = Money, Money, Money…**
  - Route import and export driven by business considerations
  - But *not* driven by technical considerations!
    Example: Slower route via peer may be preferred over faster route via provider

Where to Peer

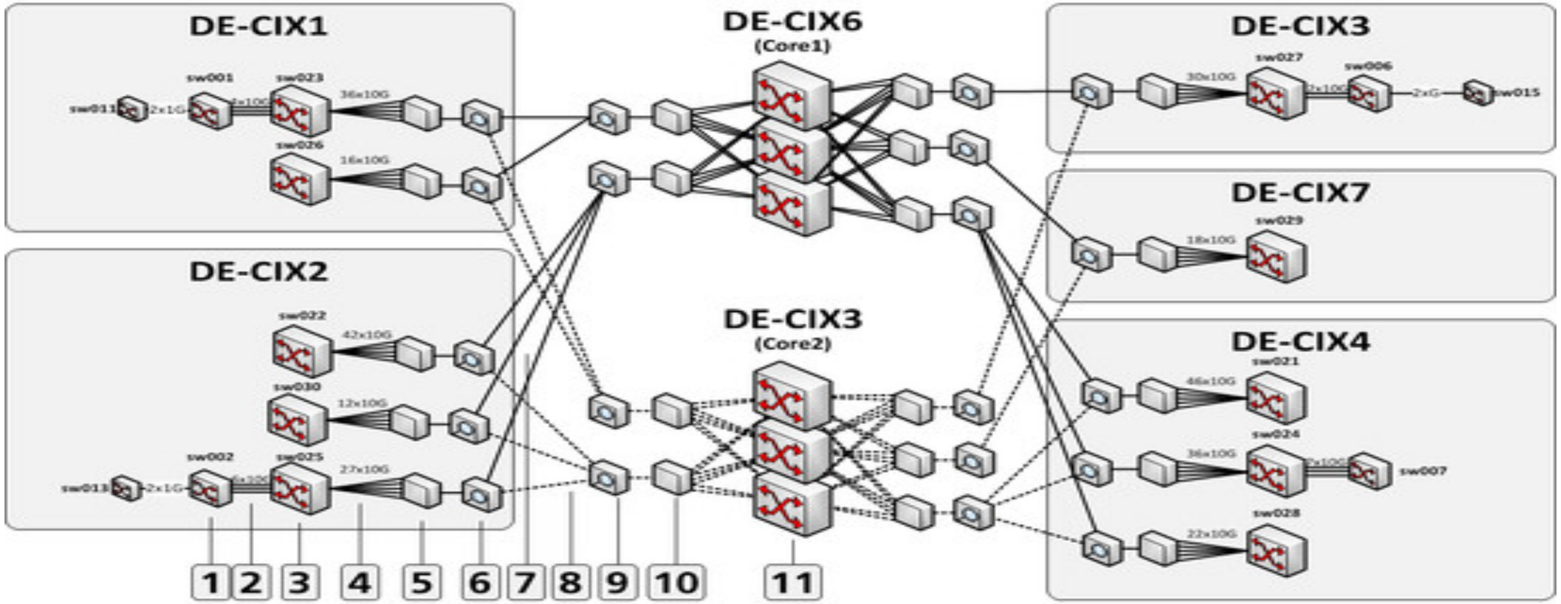
(Here: Peering = having a BGP relationship)

A) Private peering

- The obvious solution: “Let’s have a cable from your server room to our server room”

B) At public peering locations (Internet Exchange Point, IX, IXP)

- “A room full of switches that many providers connect to”
- Configure VLAN connections in switch, instead of having to put in $O(n^2)$ separate wires
- Examples:
  - DE-CIX, Frankfurt (purportedly largest in world)
  - AMS-IX, Amsterdam
  - LINX, London
  - MSK-IX, Moscow



1 Force10 Terascale E1200
2 Multiple 10G-Connections
3 Force10 Exascale E1200i
4 Multiple 10G-Connections
5 DWDM MUX 32 Channel
6 Lynx LightLeader Master Unit
7 Dark Fiber Working Line
8 Dark Fiber Protection Line
9 Lynx LightLeader Slave Unit
10 DWDM MUX 32 Channel
11 2xBrocade MLX32 and 1xForce10 Exascale 1200i per Core

- Source: de-cix.net