Master Course
Computer Networks
IN2097

Prof. Dr.-Ing. Georg Carle
Christian Grothoff, Ph.D.
Stephan Günther

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

Routing

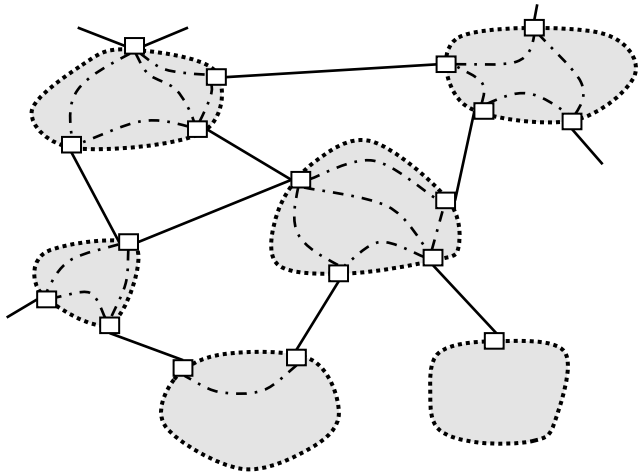
Topics

- Routing and forwarding
- Routing algorithms recapitulated
  - Link state
  - Distance Vector
  - Path Vector
- Intradomain routing protocols
  - RIP
  - OSPF
- Interdomain routing
  - Hierarchical routing
  - BGP
- Business considerations
  - Policy routing
  - Traffic engineering
- Routing security
- Multicast routing

Inconsistent Topology Information

- Typical causes (not exhaustive)
  - One router finished with calculations, another one not yet
  - Relevant information has not yet reached entire network
    - LS: Broadcasts = fast
    - DV: Receive message, calculate table, inform neighbours: slow
  - DV: Count-to-infinity problem
  - LS: Different algorithm implementations!
  - LS: Problem if there is no clear rule for handling equal-cost routes
- Possible consequences?
  - Erroneously assuming some destination is not reachable
  - Routing loops

Inter-AS Routing vs. Intra-AS Routing



- Autonomous Systems
  - World: > 37.000 Autonomous Systems
  - Europe: > 19.000 Autonomous Systems
  - Germany: > 1.200 Autonomous Systems

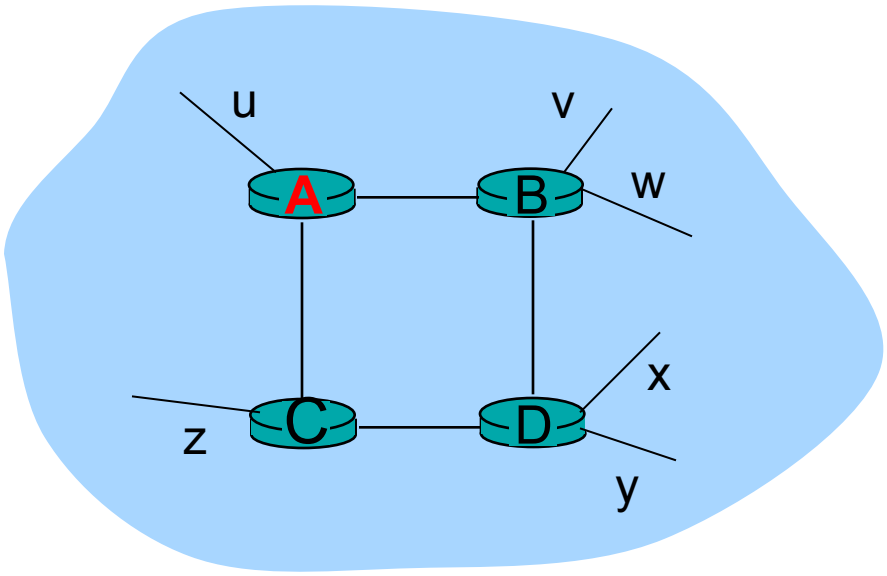
- Subsequently: closer look at Intra-AS Routing

Intra-AS Routing

- Also known as Interior Gateway Protocols (IGP)
- Most common Intra-AS routing protocols:
  - RIP: Routing Information Protocol — DV (typically small systems - lower tier or enterprise networks)
  - OSPF: Open Shortest Path First — hierarchical LS (typically medium to large systems - upper tier ISPs)
  - IS-IS: Intermediate System to Intermediate System — hierarchical LS (typically medium-sized ASes)
  - (E)IGRP: (Enhanced) Interior Gateway Routing Protocol (Cisco proprietary) — hybrid of LS and DV

RIP - Routing Information Protocol

- Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982
- Distance metric: # of hops (max = 15 hops, ∞ ≔ 16)
- Sometimes still in use by small ISPs



From router A to subnets:

destination	hops
u	1
v	2
w	2
x	3
y	3
z	2

OSPF - Open Shortest Path First

- "Open": publicly available (vs. vendor-specific, e.g., EIGRP: Cisco-proprietary)
- Uses Link State algorithm
  - LS packet dissemination (broadcasts)
  - Unidirectional edges (⇒costs may differ by direction)
  - Topology map at each node
  - Route computation using Dijkstra's algorithm
- OSPF advertisement carries one entry per neighbour router
- Advertisements disseminated to entire AS (via flooding)
  - (exception: hierarchical OSPF, see subsequent slides)
  - carried in OSPF messages directly over IP (no TCP or UDP)

OSPF "Advanced" Features (not in, e.g., RIP)

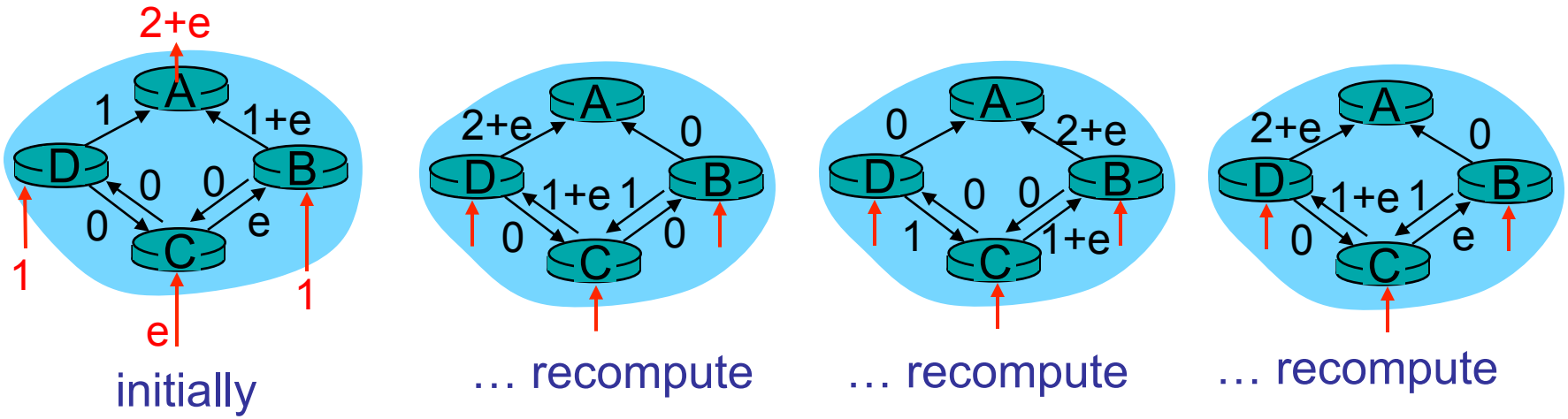
- Security: all OSPF messages authenticated (to prevent malicious intrusion)
- Multiple same-cost paths allowed (only one path in RIP): *ECMP* (equal-cost multipath) for link load balancing
- For each link, multiple cost metrics for different Type of Service (TOS):
  e.g., satellite link cost set to "low" for best effort, but to "high" for real-time traffic (e.g. for telephony traffic)
- Integration of multicast support:
  - Multicast OSPF (MOSPF) - RFC1584
  - Uses same topology data base as OSPF
    → less routing protocol traffic
- Hierarchical OSPF in large domains
  - Significantly reduces number of broadcast messages

OSPF (Open Shortest Path First)

- RFC 2328: OSPF v2, 244 pages
- Link advertisements broadcasts
  - after state change (cost, up/down status)
  - periodically (at least every 30 minutes)
- Authentication of advertisements
  - different authentication procedure possible for each subnet
  - auth type field and 64-bit auth field in OSPF packet header
  - message digest appended to OSPF packet header
- OSPF Protocol does not specify policy for how to set link costs
  - local decision (AS administrator)
  - OSPF is just the protocol for least-cost-routing
  - *traffic engineering with OSPF*: setting link costs such that OSPF routing results in desired routing of flows

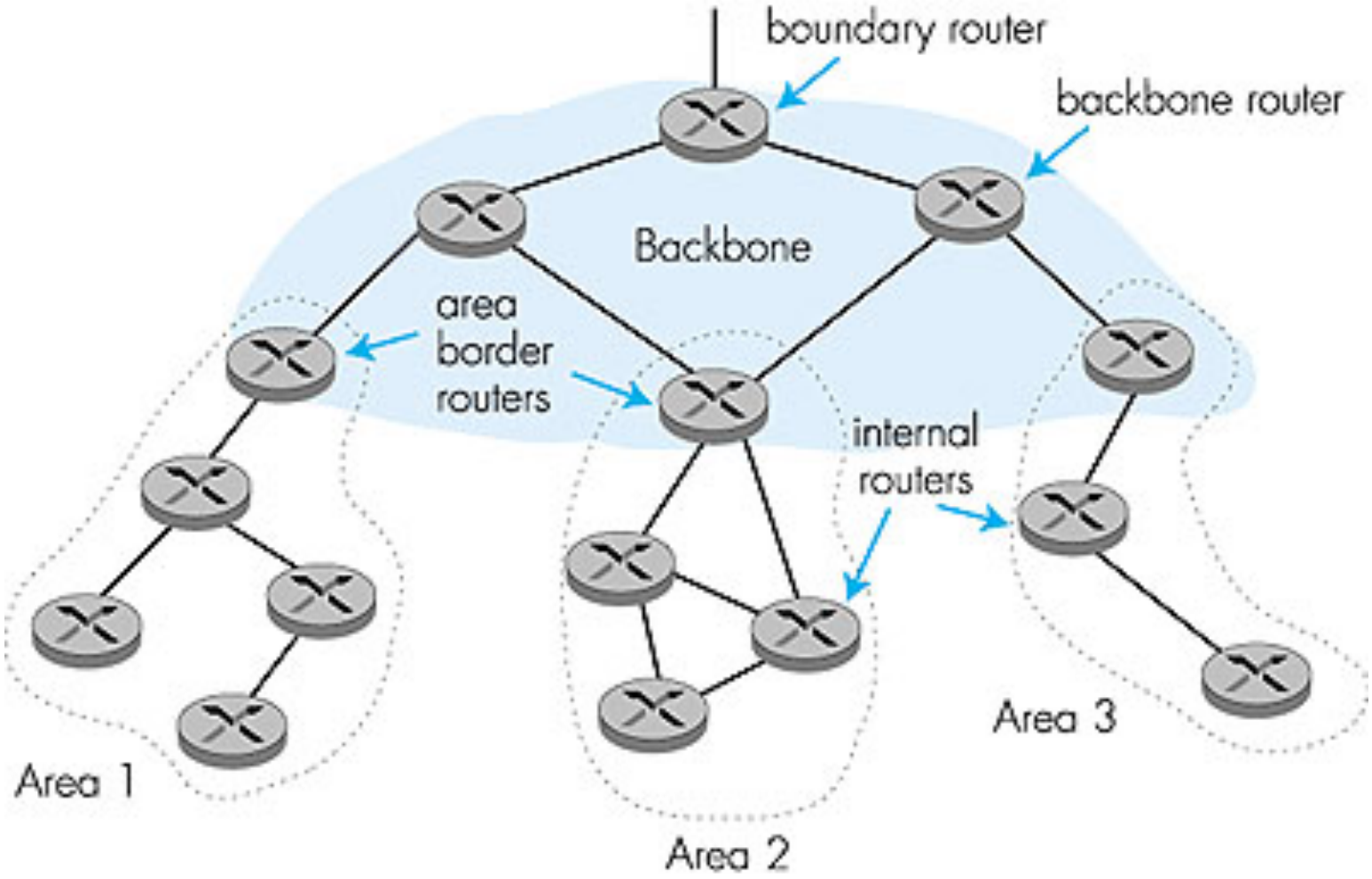
Dynamic (Congestion-Sensitive) Routing?

- Congestion-sensitive routing may lead to oscillations
- e.g., link cost = amount of carried traffic
  - implication: link costs not necessarily symmetric



- Why is this a bad thing?
  - Possibly sub-optimal choice of paths (as in example above)
  - Inconsistent topology information during convergence

Hierarchical OSPF



Hierarchical OSPF

- OSPF *can* create a two-level hierarchy
  - (similar, but not identical to inter-AS and intra-AS routing within an AS)
- Two levels: local *areas* and the *backbone*
  - Link-state advertisements only within local area
  - Each node has detailed area topology;
    but only knows coarse direction to networks in other areas (shortest path to border router)
- *Area border routers:* "summarize" distances to networks in own area; advertise distances to other Area Border and Boundary routers
- *Backbone routers:* run OSPF routing limited to backbone
- *Boundary routers:* connect to other ASes
  - "The outside world" ≈ another area

Hierarchical Routing in the Internet

simple viewpoint

- all routers identical
- network “flat”

vs. reality:

Scale = billions of destinations:

- Cannot store all destinations in routing tables
- Routing table exchange would swamp links
- Thousands of OSPF Areas? Would not scale!

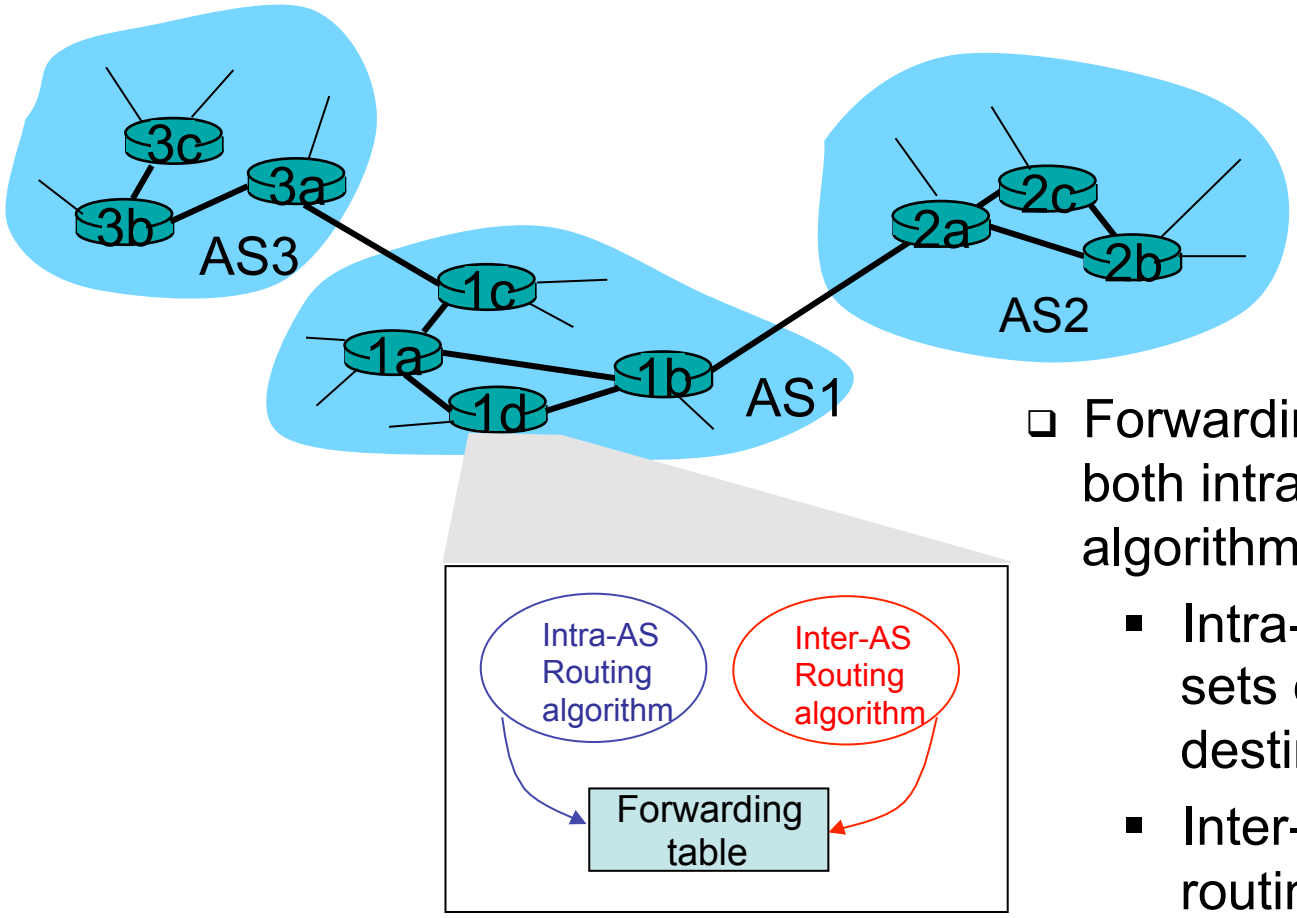
Administrative autonomy

- Internet = network of networks
- Each network admin may want to control routing in its own network — no central administration!

Hierarchical Routing

- Aggregate routers into regions called "autonomous systems" (short: AS; plural: ASes)
  - One AS ≈ one ISP / university
- Routers in same AS run same routing protocol
  - = "intra-AS" routing protocol (also called "intra-domain")
  - 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 "inter-domain")
  - Warning: *Non-gateway routers* need to know about *inter-AS* routing as well!

Interconnected ASes



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

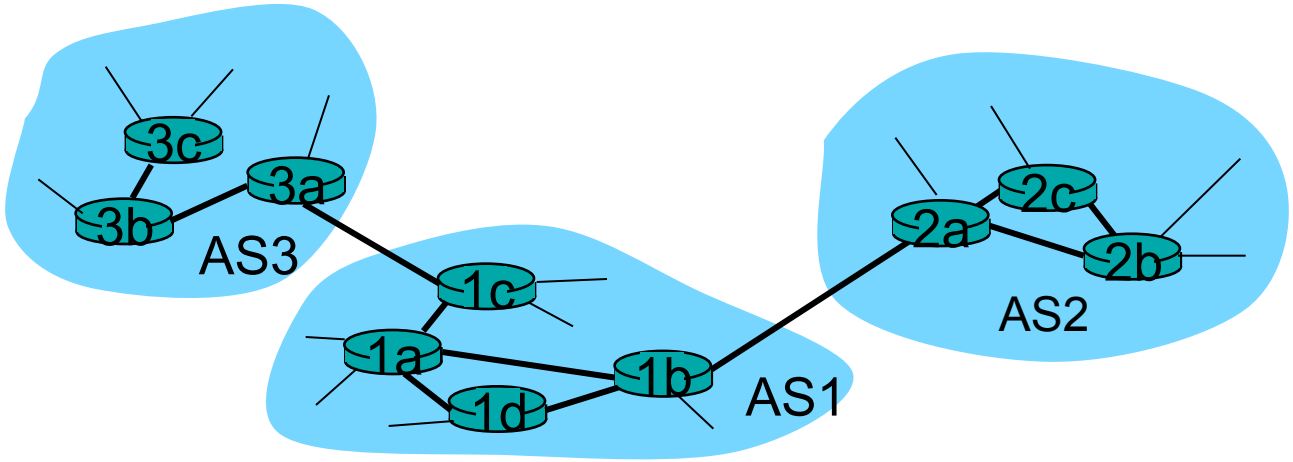
Inter-AS Tasks

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

AS1 must:

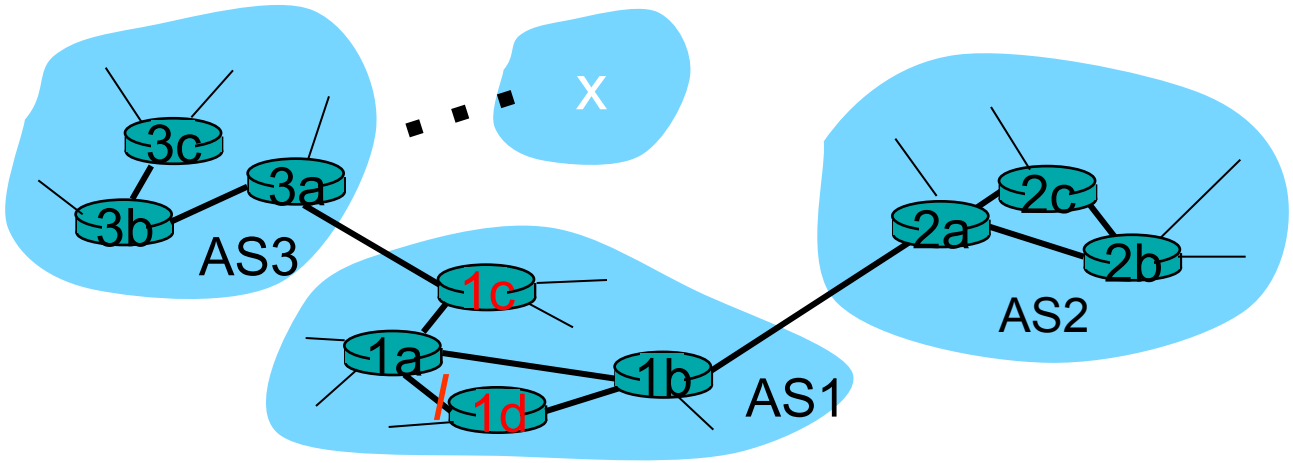
1. learn which destinations are reachable through AS2, which through AS3
2. propagate this reachability info *to all* routers in AS1 (i.e., not just the gateway routers)

Job of inter-AS routing!



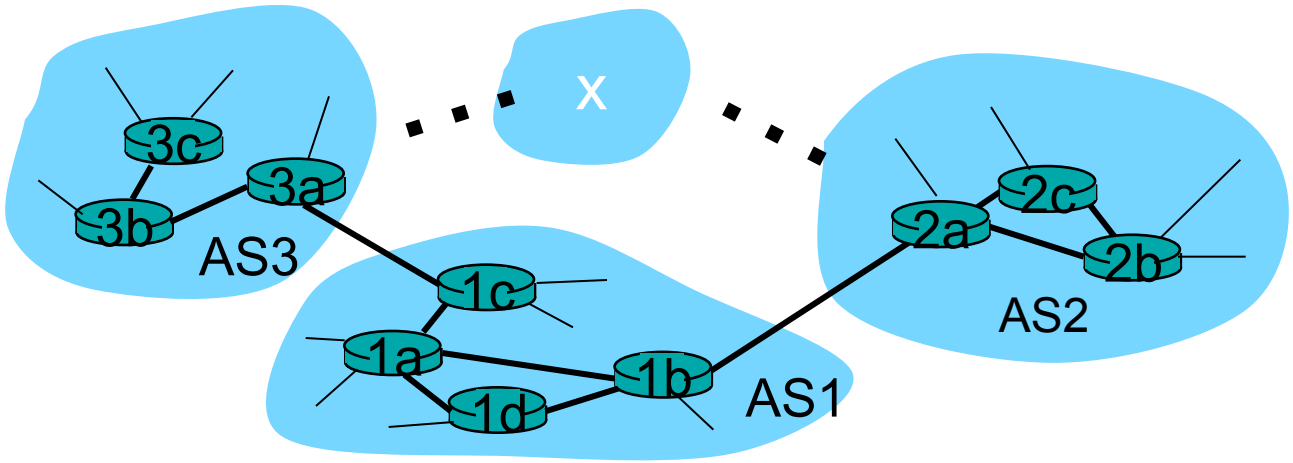
Example: Setting forwarding table in a router

- 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!

Path Vector protocols

- Problem with Distance Vector protocol:
  Path cost is “anonymous” single number; does not contain any topology information
- Path Vector protocol:
  - For each destination, advertise entire path (=sequence of node identifiers) to neighbours
  - Cost calculation can be done by looking at path
    - E.g., count number of hops on the path
  - Easy loop detection: Does my node ID already appear in the path?
- Not used very often
  - only in BGP …
  - … and BGP is much more complex than just paths

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 neighbouring 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)

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)

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

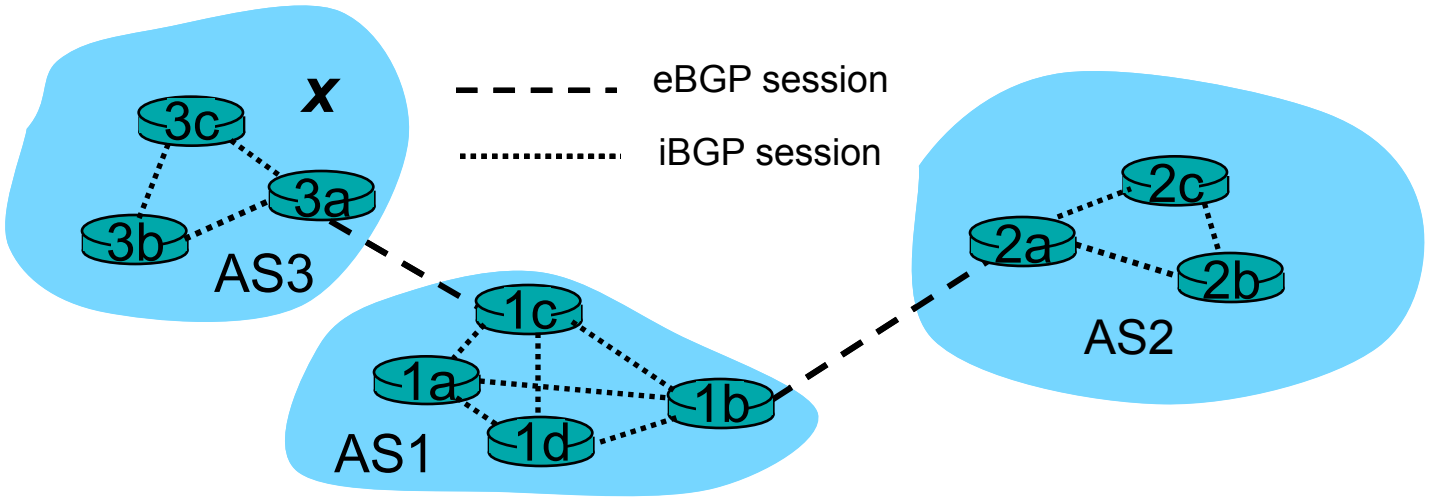
BGP Updates

- 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 Discriminators - e.g. to announce which of several links are preferred for inbound traffic
  - Community: attribute tags applied to prefixes to achieve some common goal on how prefixes are to be treated, e.g. geographic or peer type restrictions (RFC 1997).
    - community attribute is transitive, but communities applied by customer rarely propagated outside next-hop AS

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

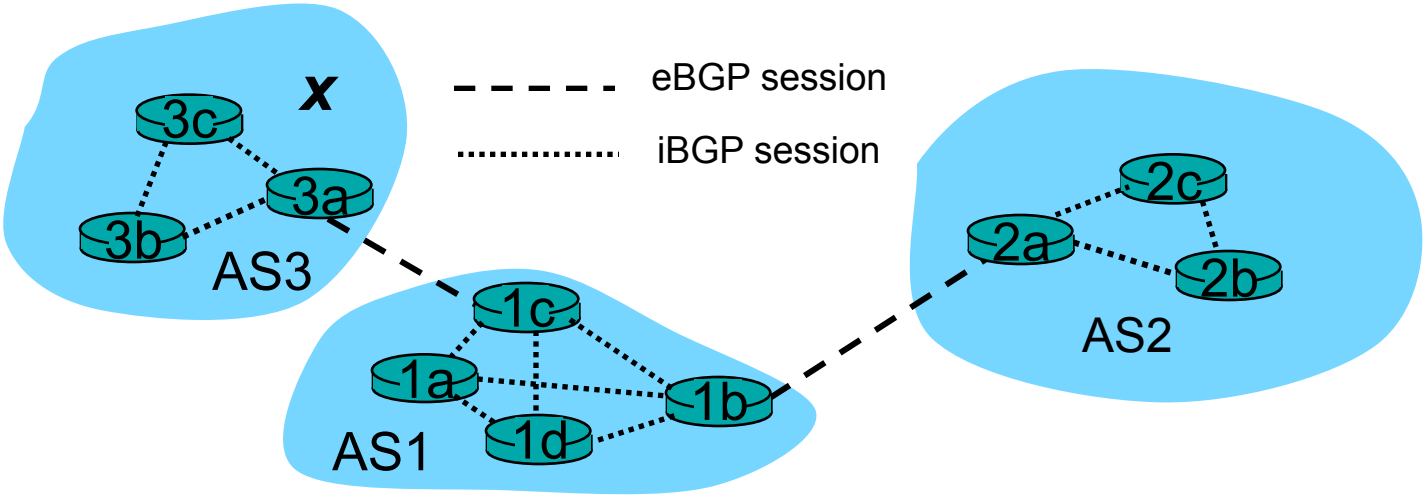
eBGP and iBGP

- 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)
  - full IBGP mesh, or route reflectors, or confederations
- No different protocols—just slightly different configurations!



Distributing Reachability Information

- 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:
  - TUM-I8-AS = AS56357
  - Leibniz-Rechenzentrum = AS12816
  - Deutsche Telekom = AS3320
  - AT&T = AS7018, AS7132, AS2685, AS2686, AS2687
- ASNs used to be 16bit, but can be 32bit nowadays
  - May have problems with 32bit ASNs on very old routers
- ASN assignment: similar to IP address space
  - ASN space administered IANA
  - Local registrars, e.g., RIPE NCC in Europe

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

Originator:
The AS that "owns" 10.11.128.0/17

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

How the update travelled

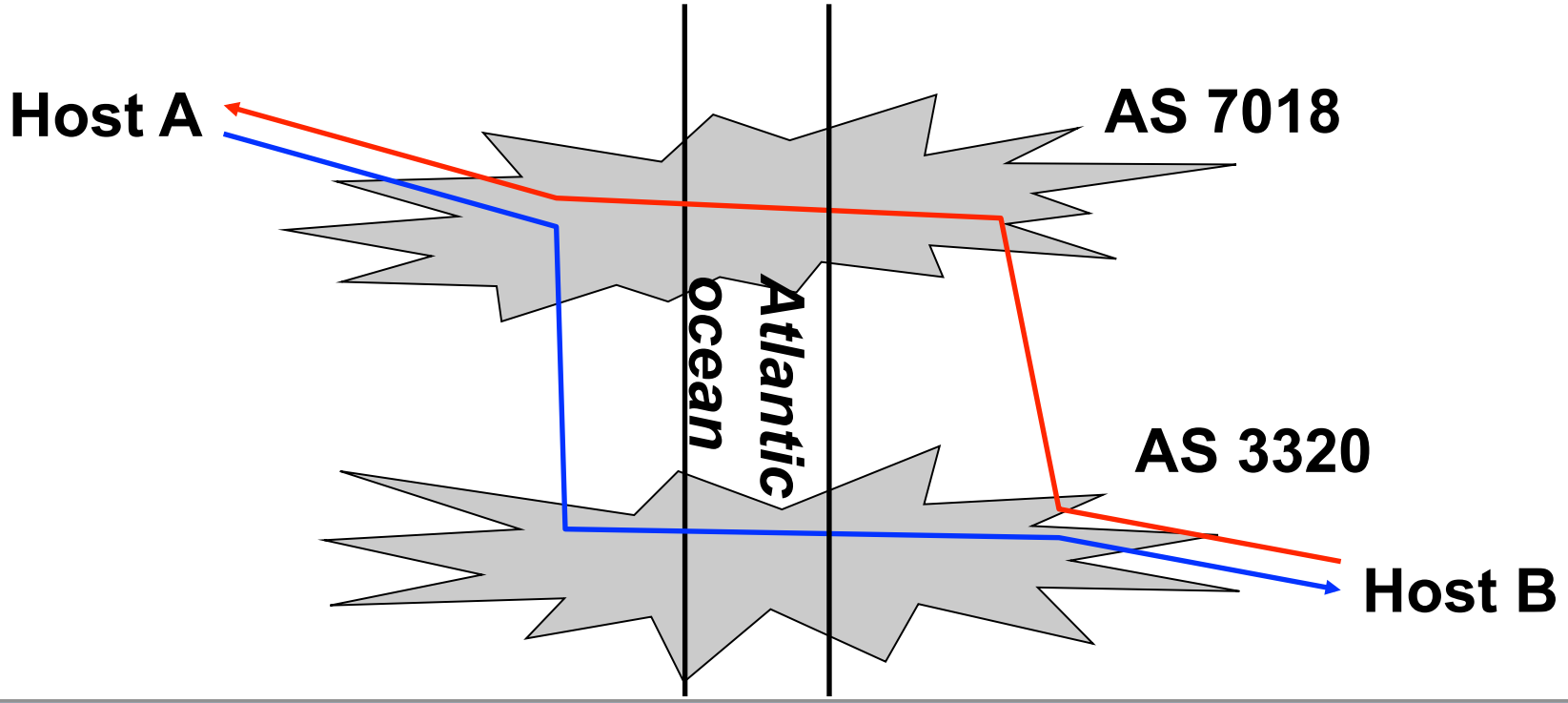
How the IP packets will be forwarded (if this route gets chosen)

BGP Route Selection

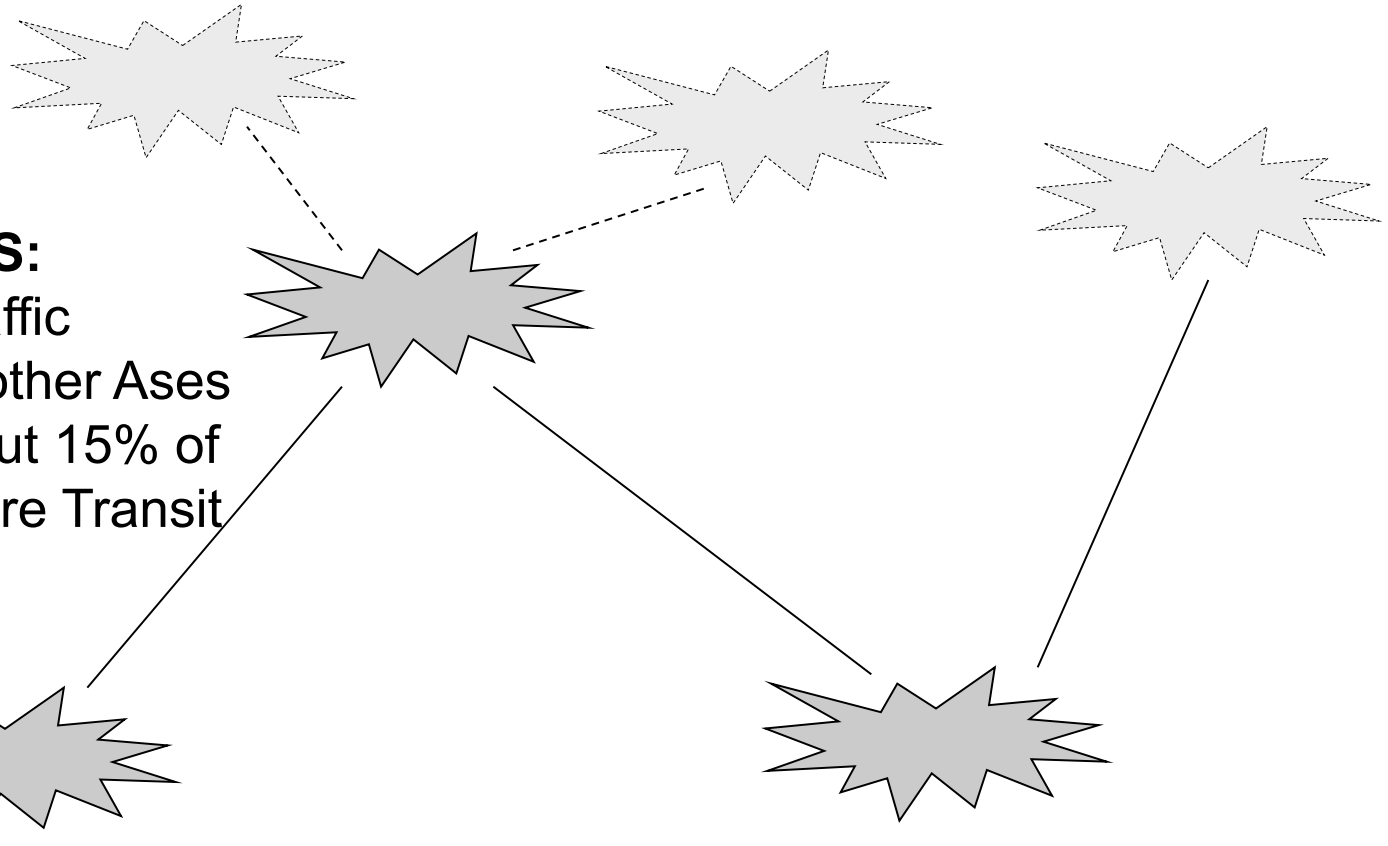
- Router may learn about more than 1 route to some prefix
  ⇒ Router must select the best one among these
- Elimination rules (**simplified**):
  0. WEIGHT: local to the router (i.e. not transmitted by BGP)
  1. Local preference value attribute: policy decision
     - if there are several routes, the one with the highest LOCAL_PREFERENCE is chosen
  2. Shortest AS-PATH
  3. Closest NEXT-HOP router
     - hot potato routing (→ next slide)
  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

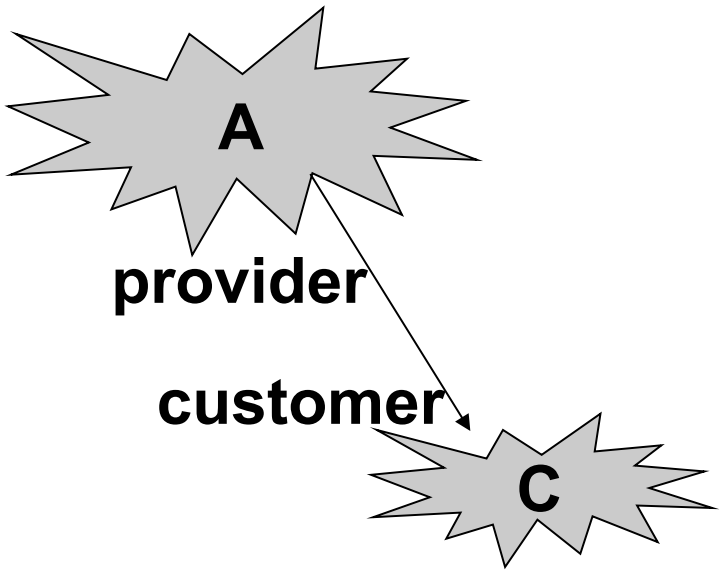
- Internet = network of networks (ASes)
  - Many thousands of ASes
  - Not every network connected to every other network
  - BGP used for routing between ASes
- Differences in economical power/importance
  - Some ASes huge, intercontinental (AT&T, Cable&Wireless)
  - Some ASes small / local (e.g., München: 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)

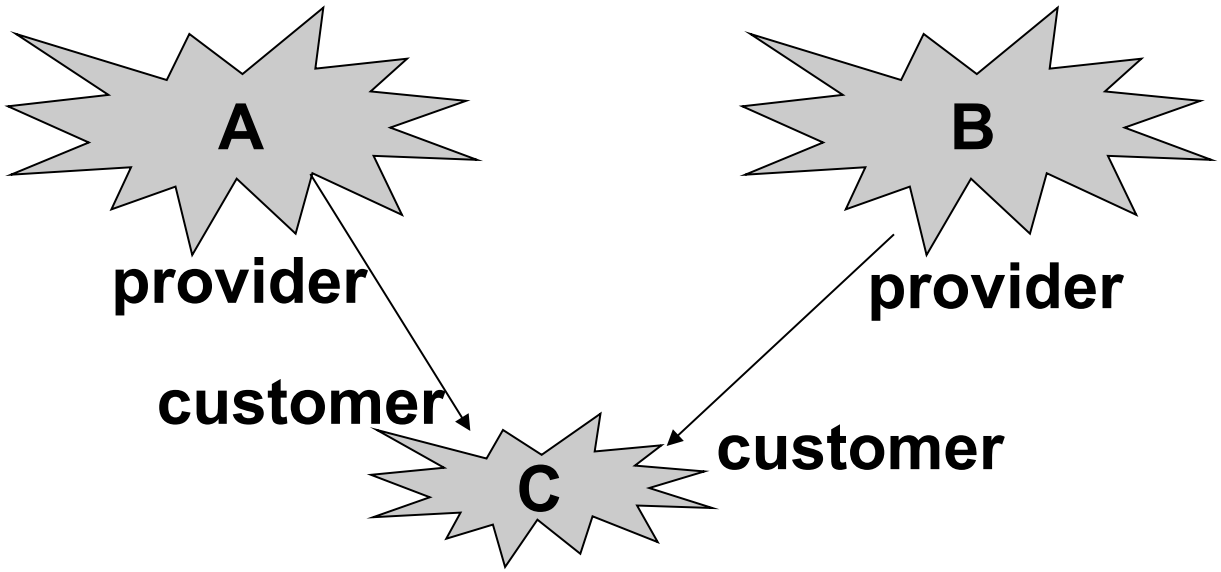
Business and policy routing (2)

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



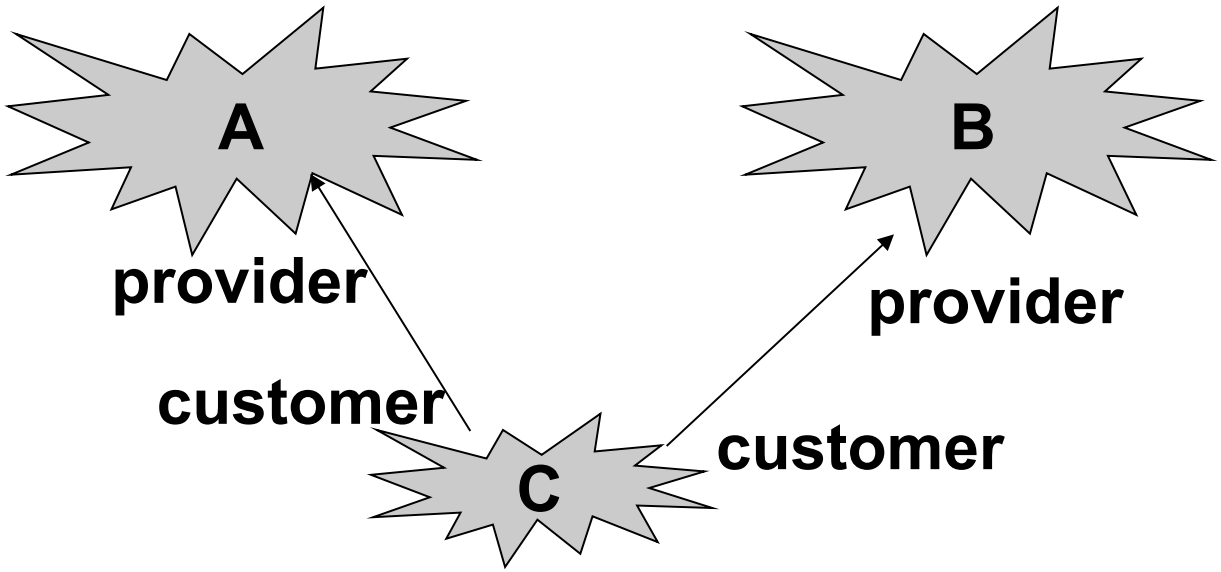
Business and policy routing (3)

- 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 / …



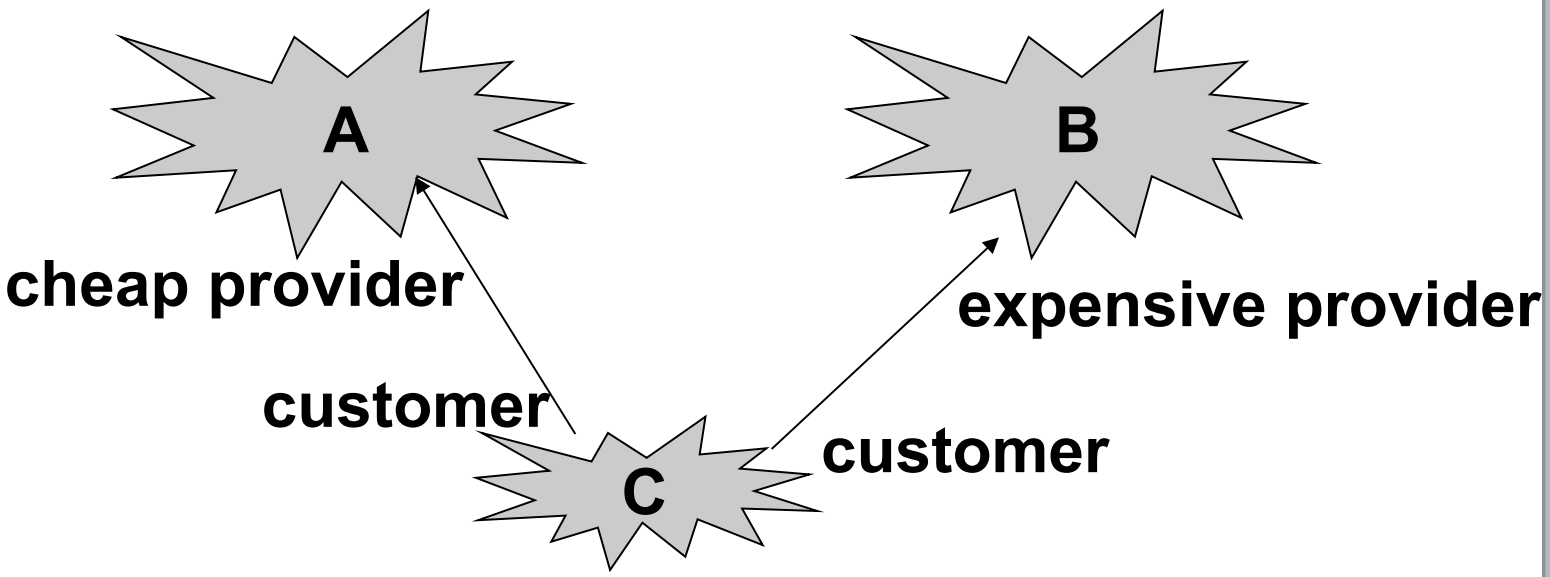
Business and policy routing (4)

- 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_PREFERENCE* 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