Unicast Routing

TCP/IP class

Routing Protocols

- intro
- RIP and son of RIP
- OSPF
- BGP
- odd bodkins
  - NAT

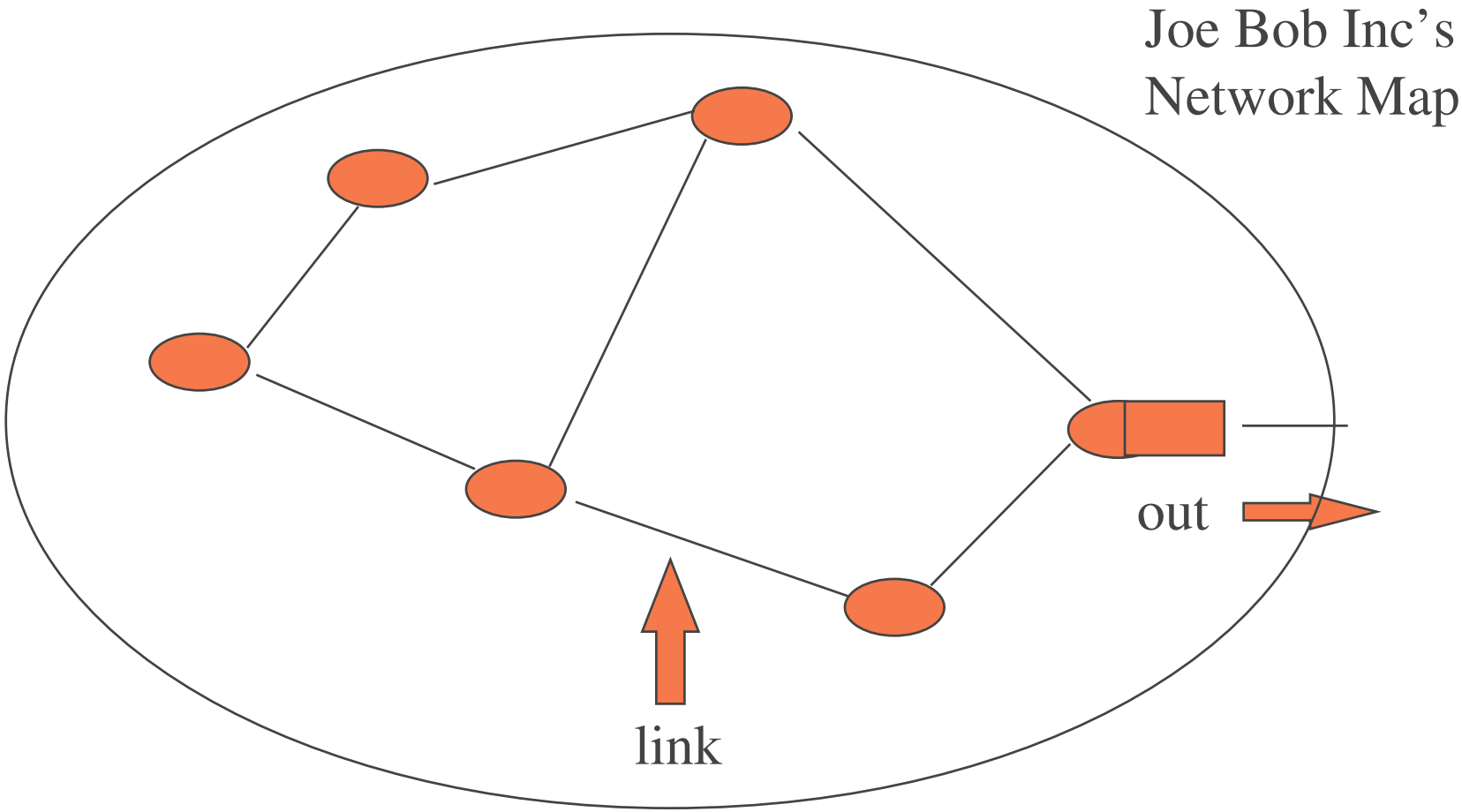
divide routing world into 3 parts

topology	IETF	ISO/OSI
same “link” or wire	none, intra-link?	none, intra-link?
enterprise or campus	Interior Gateway Protocol - IGP	intra-domain routing protocol
between enterprises	Exterior Gateway Protocol - EGP	inter-domain

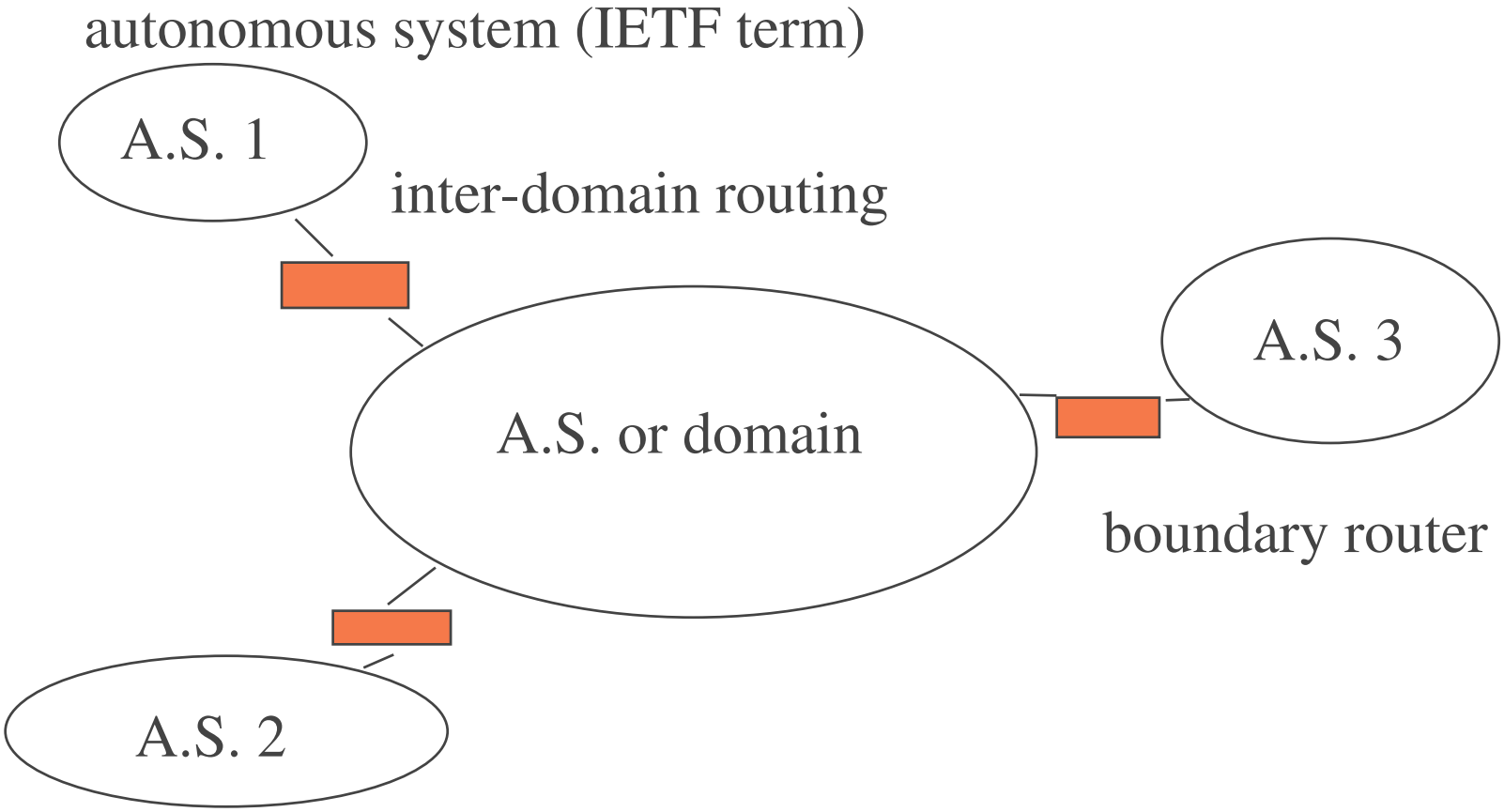
protocols acc. to topology

topology	IETF	ISO/OSI
intra-link	ARP	ES-IS
intra-domain	RIP, RIP(2), OSPF	IS-IS
inter-domain	EGP, BGP(4)	IDRP

the Interior - RIP or OSPF



Exterior - Between Domains - BGP



Routing Information Protocol

- BSD app based on XNS (Xerox) version, Netware RIP is similar too (surprise)
- done first and RFC 1058 (1988) later created
- in widespread use for at least two reasons
  - widely available, came with that there Sun
  - # routed & is all you need to do
- BSD routed and Cornell gated support it

RIP details

- messages carried in UDP datagrams, send/recv on port 520
- broadcast every 30 seconds, routing table as pairs of (to net, hop count)
- triggered update sent if metric (hop count) changes (only relevant info)
- hop count, direct connect == 1, network one router away is 2 hops away
- new route with shorter hop count replaces older route
- on init, router requests route table from neighbors

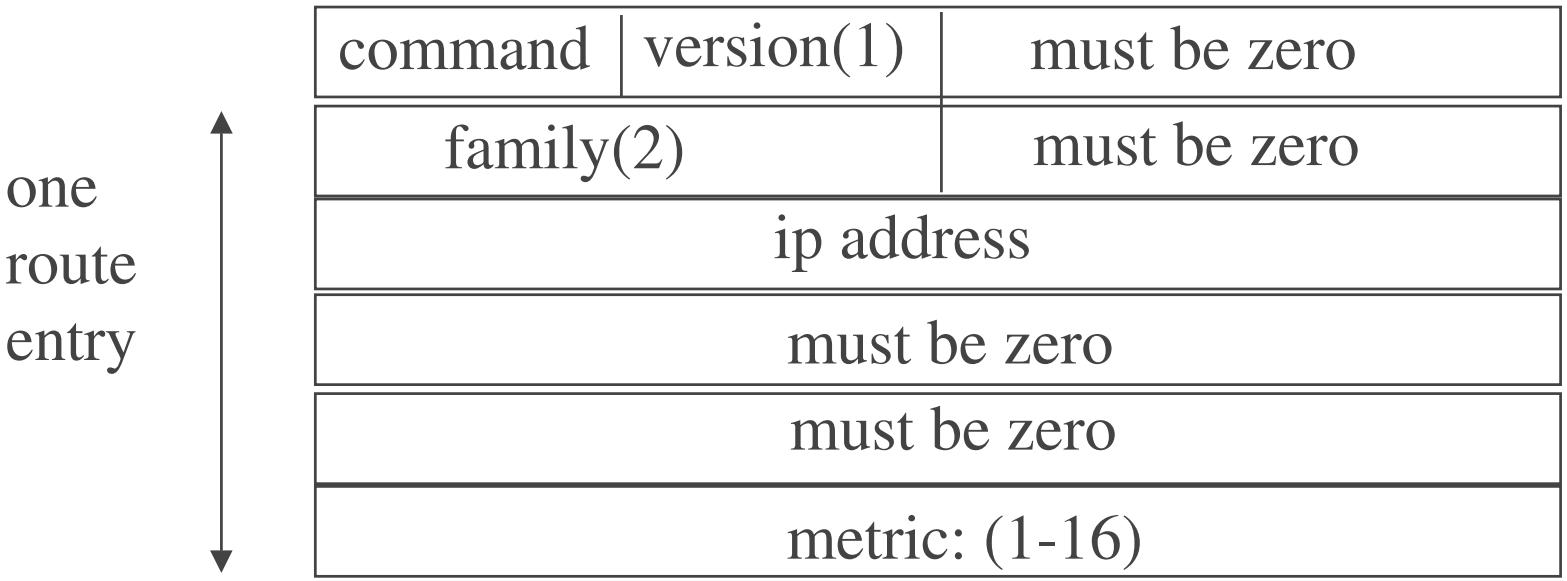
more RIP details

- when routing response receiving, routing table is updated (metrics aren't typically displayed in netstat -rn unfortunately)
- route has timeout. 3 minutes, no new info, then mark with metric=16, one minute later delete (holddown so the fact that route is gone is propagated)
- infinity == 16, RIP can suffer count to infinity
- default route is route to 0.0.0.0
- routers are "active", hosts are "passive", determined by whether or not system > 1 i/f (can set by hand)

to RIP or not to RIP?

- pros
  - simple, stupid...
- cons
  - no understanding of subnetting; e.g.,
    - » 121.12.3.127 could be a host or a subnet paired with 121.12.0.0 leads rip to think what?
  - convergence is slower (minutes sometimes) AND
  - not as scalable as OSPF - can't aggregate as well
    - » hop count max is small (not really important)
    - » can't deal with different link types

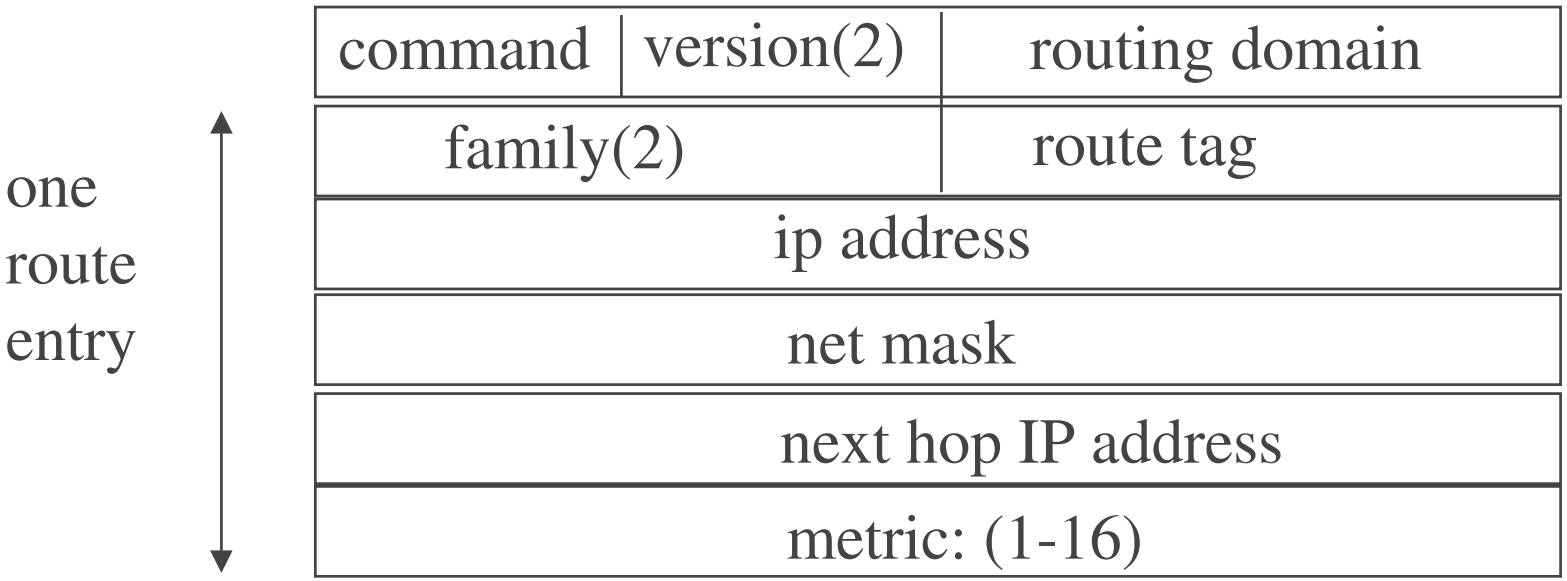
RIP(1) header



up to 24 more routes, 25 routes max (< 512)

note: command: 1, request; 2, response

RIP(2) header



up to 24 more routes, 25 routes max (< 512)

RIP-2

- RFC 1388 (1993)
- zero fields cleverly used, should interoperate if RIP(1) ignores fields
- version is 2
- routing domain can be used to allow more than one RIP domain on a campus; more than one routed on a system
- route tag - AS number, communicate boundary info
- subnet mask - for CIDR, route == (ip, net mask)
- next hop, ip address for VIA part of route (as opposed to getting it from IP src)

RIP-2

- clear-text password
  - shared-secret e.g., with MD5 can exist though
- can use multicasting as opposed to broadcast, thus hosts that
  - "don't give a RIP(2)" can ignore it

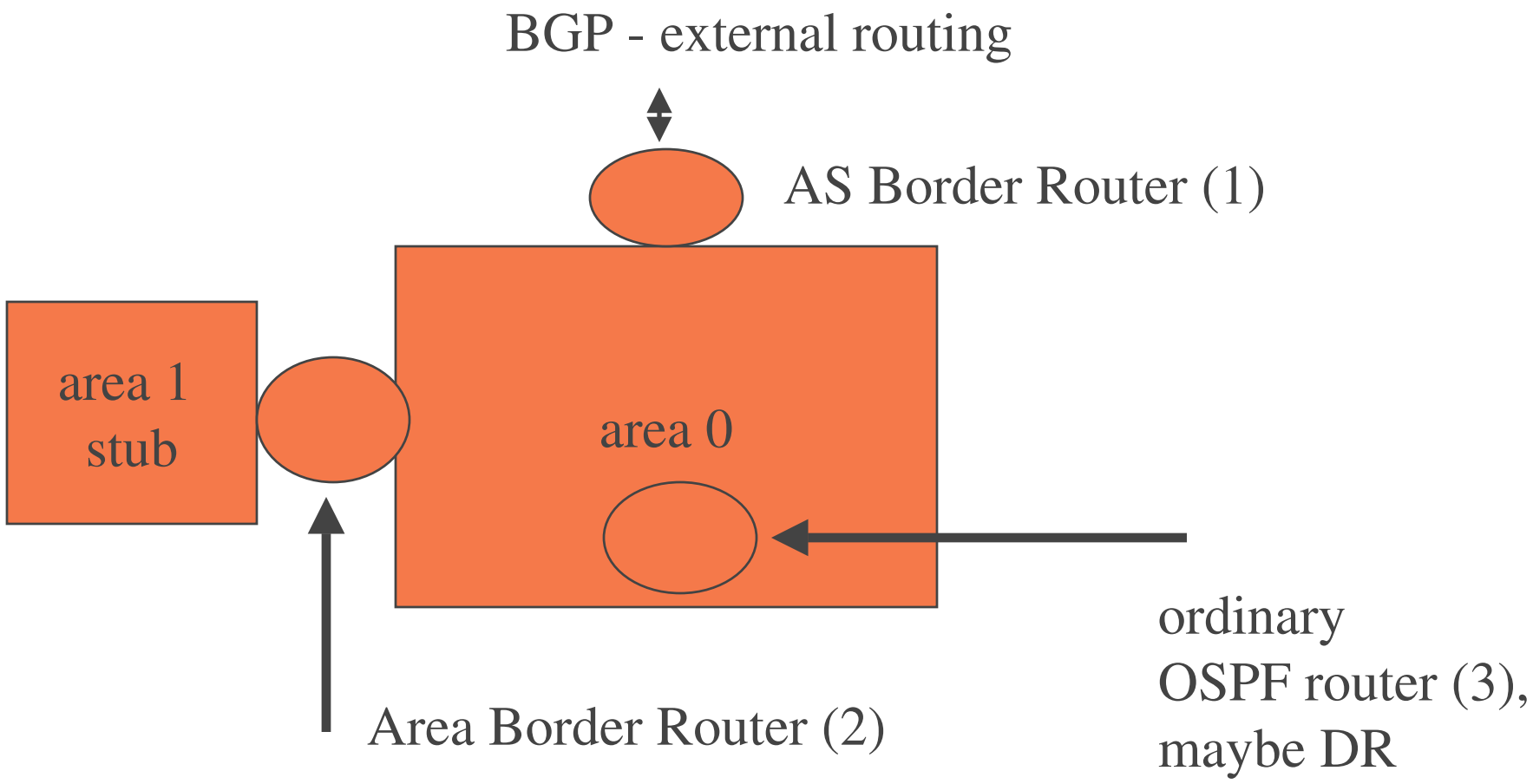
OSPF - Open Shortest Path First

- OSPF version 2, in RFC 1247 (1991)
- **link-state** protocol as contrasted with RIP
- OSPF uses IP direct, not on top of UDP, proto = 89
- OSPF has **backbone routers** (top level/L2) and (lower level/L1) **internal routers**
- supports AREA notion**,** **backbone** router can summarize IP addresses in area, report summary to other backbone routers, and leak that info into area so that internal routers can optimize their routes
- uses multicast as opposed to broadcast

OSPF

- routers can do **load balancing** if more than one path and metric is the same
  - equal-cost multi-path routing
- metrics are in theory dimensionless, in reality: link speed (ethernet is 1000/100/10 ...)
- one router on link plays the LSP game, **designated router**, has election algorithm
- supports subnets (CIDR), host route has mask of all 1's, default all 0's

OSPF router types



router functions

- ABSR - runs BGP/OSPF
  - decides how much external BGP routing info to interject into A.S. (and vice versa)
- Border Router - aggregates area external and internal routes and injects into other area (summaries)
- DRs and non-DRs, participate in OSPF within an area

OSPF - 3-sub-protocols

- hello
  - routers on same link exchange link info
  - elect DR - designated router
- exchange
  - bringing up adjacencies
  - routers at (re)boot exchange Link-State tables
- update
  - flooding of link-state change/includes ACK

Link-State record types (5)

- router LSP - sent by routers within AREA
  - describes links and associated costs (metrics)
- network LSP - sent by DR, within AREA only
  - describes other routers on link
- IP network summary - Area Border Routers send across areas
  - aggregation of one area to another

LSP types, cont.

- border router summary - ABRs send
  - describes path to ASBR
- external - ABSRs send IN
  - describes path to outside world
- note 1st two describe AREA setup
- last 3 describe into/out of AREAs/A.S. and
  - include aggregation

BGP - Border Gateway Protocol

- bind A.S. or domains together (Layer 3?). A.S. is 16 bit number allocated by regionals (e.g., ARIN in US)
- replaced EGP, see RFC 1457 and possibly newer versions
- BGP uses TCP to communicate
  - reliable
  - can tunnel across a domain
- distance vector protocol. route == series of A.S. numbers, since route is enumerated, can detect loops
- route update = To X, AS #1, AS #2, etc.

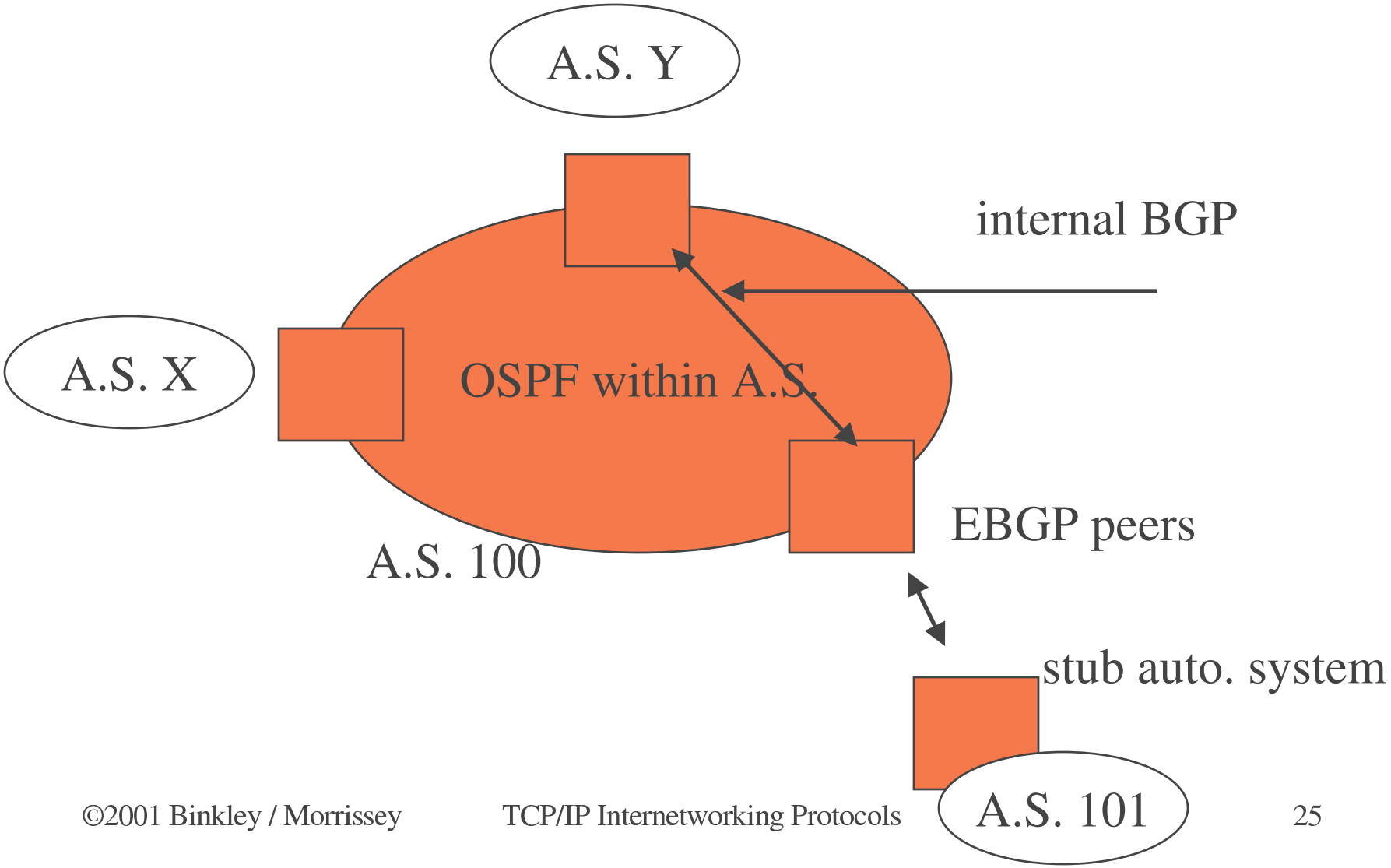
BGP continued

- AS is either:
  - stub: only one way in/out
  - transit: in the middle of stub A.S.
  - multi-homed: more than one way out but refuses to do transit work
- routing can be policy-based but is typically hop based
- policies are determined by admin and put in config files

A.S. must be obtained from

- relevant authorities (www.arin.net in US)
- view IP address as two tuple (A.S., IP)
- BGP routes in implicit sense via A.S. tuple
- BUT explicit still routes to IP net
- hope network is highly AGGREGATED
- could be serial host (worst case)

transit A.S. POV



BGP protocol types

- hello
  - can take MD5 checksum (authentication) but not in use (yet)
- notification (error)
  - loop detected (example)
  - failure in TCP state machine
- update (or withdrawal)
  - route change

two types of BGP

- external, typically TCP on 2 directly connected links between two A.S.
- internal - cross BGP routers across transit A.S. (normally), may be multi-hop
  - internal BGP routers must be fully meshed; i.e., should have 1-1 connection between all BGP routers
  - OSPF must converge internally before BGP, else potential of **BLACK HOLE**

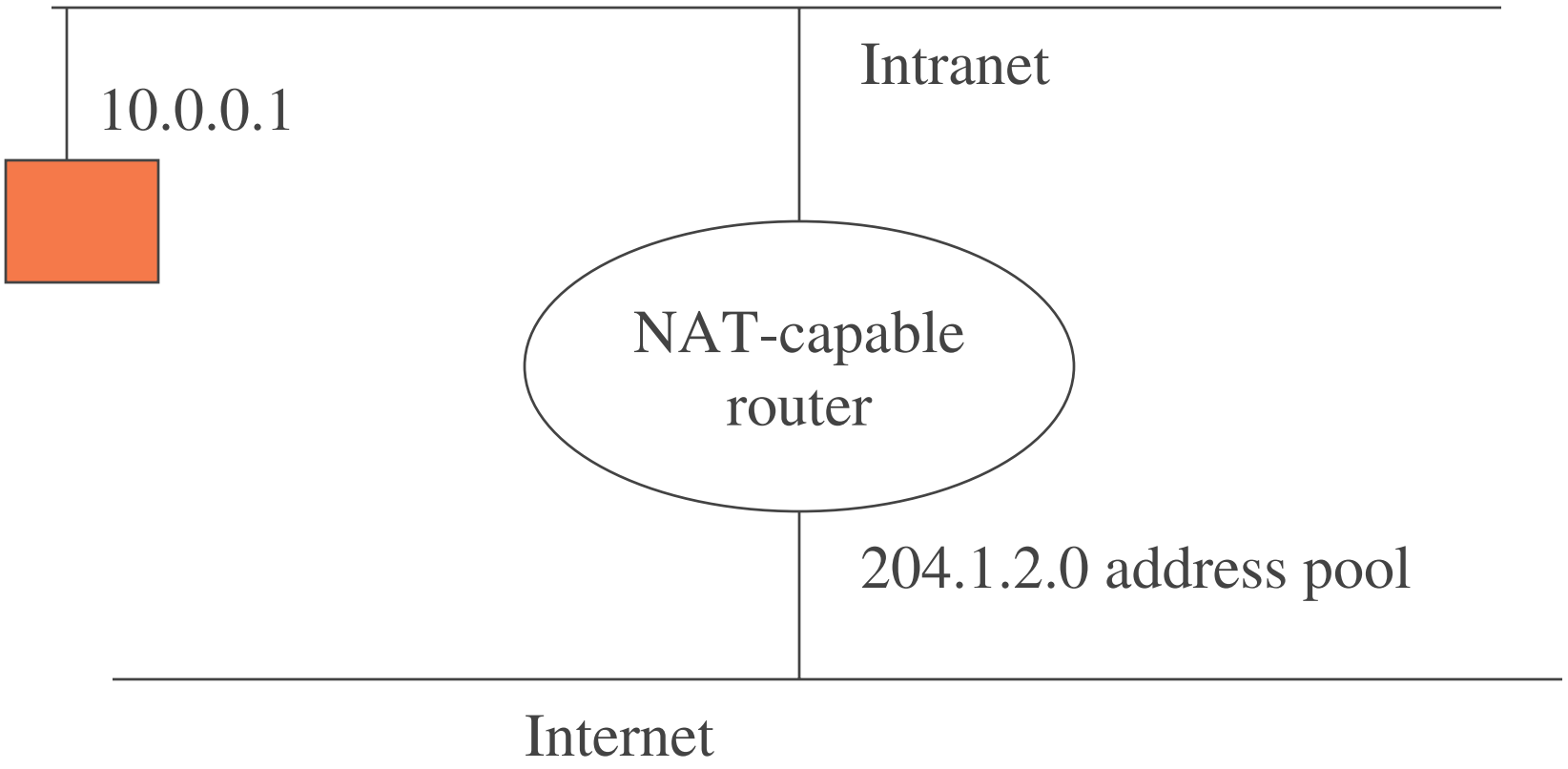
Network Address Translation

- RFC 1918 specifies a set of internal-only “intranet” addresses in range:
  - class A 10.0.0.0
  - class B 172.16.0.0 .. 172.31.255.255
  - class C 192.168.0.0 .. 192.168.255.255
- NAT idea: internal systems use private IP address - somehow mapped at router to “real” ip address

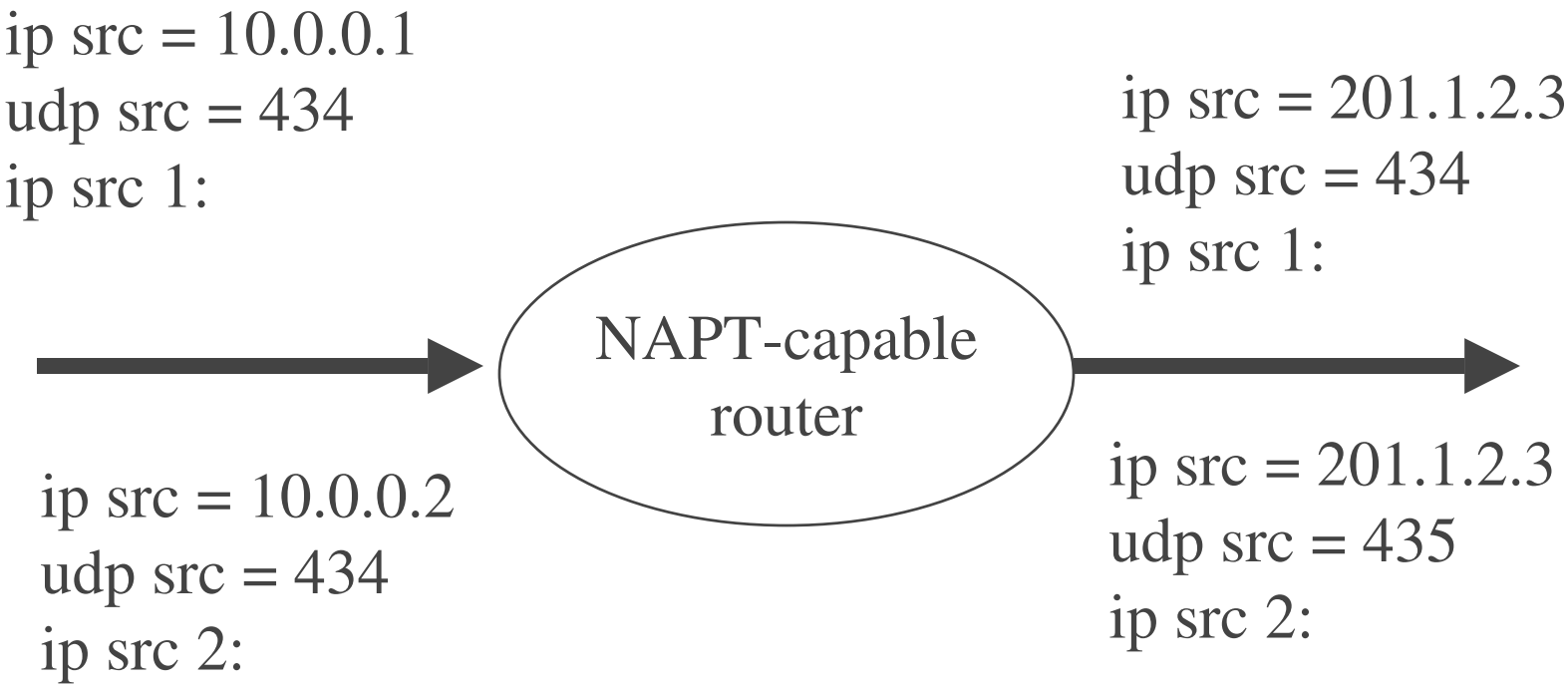
two possible configurations

- you have lots of hosts, use 10.0.0.0 internally, ip address mapping only used,
  - but one class C address externally 204.1.2.0
  - possibly this limits your external tcp connections of course
  - e.g., 10.0.0.1 is mapped to 204.1.2.1 during a tcp connect
- NAT with ports (NAPT). One external IP address. Add tcp/udp port space to make unique mapping.

NAT picture


Intranet
10.0.0.1
NAT-capable
router
204.1.2.0 address pool
Internet

NAT with ports picture - Mapping Example



what observations can you make about how the NAT box must work?

pros/cons

- pros
  - may allow administrative domain to shield all hosts from address change needed by ISP switch
  - may have security function/s
    - » outside can't see inside or can't talk to inside
    - » inside IP address changes from one connection to next therefore privacy function
  - can conserve IP address space or better utilize it
  - may have way to map one virtual address to N real addresses and get a load balancing function for server

con:

- traditional: loss of end-end connectivity
  - breaks end to end model
  - MIP won't work
  - IPSEC won't work
  - FTP ... needs passive mode, because
    - » of IP addresses in L4
- DNS functionality and tie-in?