BGP

and the

physical infrastructure

of the

Internet

Iljitsch van Beijnum

Groningen/Amsterdam 23/24 april 2003

In telephony land

- 1903: already 3.2 million subscribers in the US! Distance limited to around 2000 km (1250 mi)
- 1915: vacuum tube amplifiers, now also long distance over phone wires on poles
- 1950s: long distance circuits multiplexed over microwave radio and coaxial cable (analog!)
- Starting in the early 1960s: transistor amplifiers and digital communication

Digital telephony (or data)

- DS0 or "B channel": 64 kbps, 1 conversation, 8 kHz sampling rate and 8 bits: not CD quality!
- T1 of DS1: 1544 kbps, 24 DS0s (but usually one of those is the "D channel" used for signalling)
- T2: huh? (6312 kbps, rarely used)
- T3 or DS3: 44736 kbps, 28 T1s or 672 DS0s

European standards

- DS0: 64 kbps, no difference
- E1: 2048 kbps, 31 DS0s or 30B+D
- E2: huh? (8448 kbps, also rare)
- E3: 34368 kbps, 16 E1s, 480/496 DS0s
- E4: 139264 kbps, 4 E3s

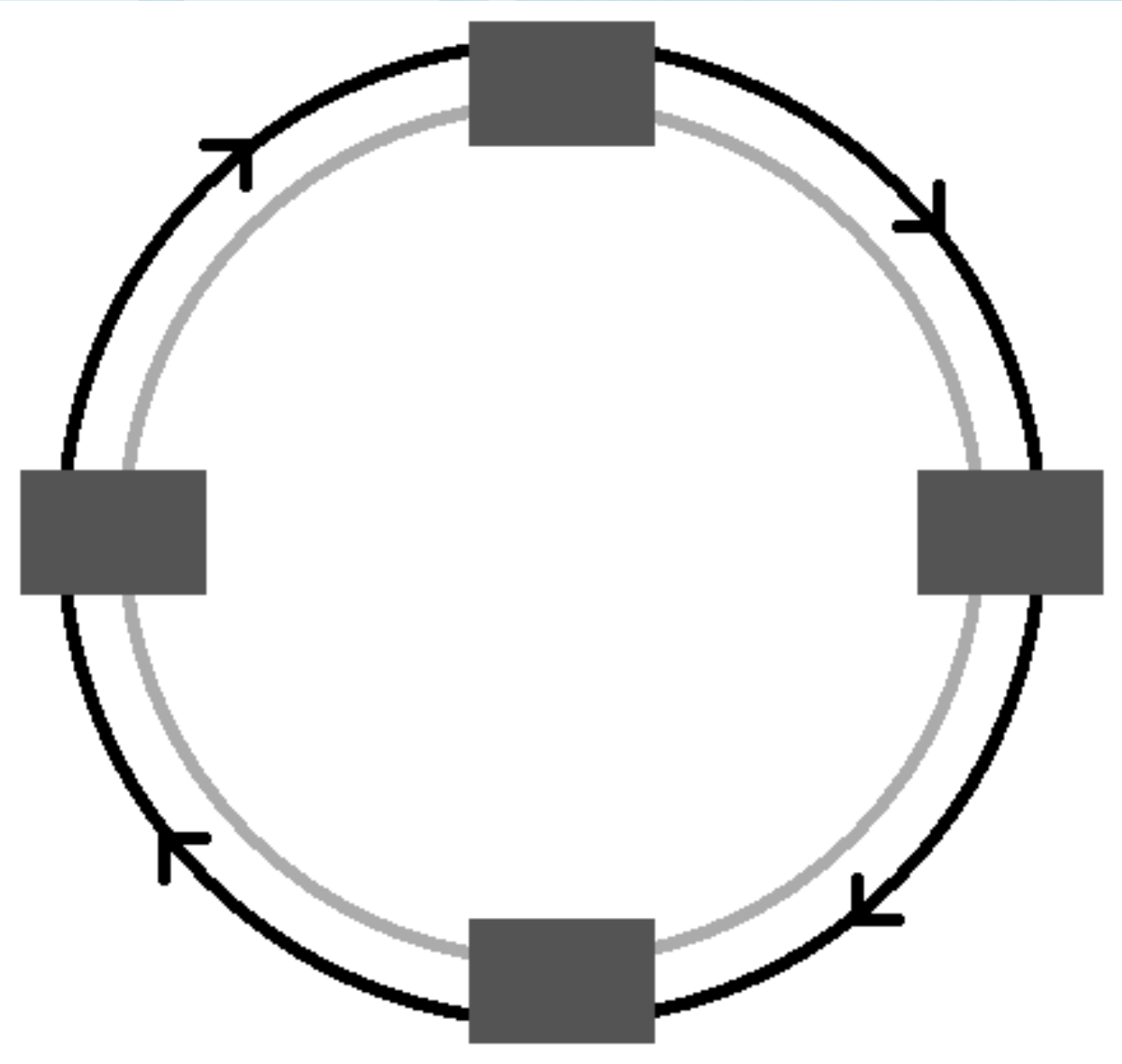
Cables

- Phone wire: unshielded twisted pair (UTP) copper wire (analog)
- T1/E1: shielded copper wire (coaxial) or UTP
- T3/E3: coaxial
- Higher speeds: SONET/SDH 155, 622, 2488 or 9952 Mbps over optical (glass) fiber

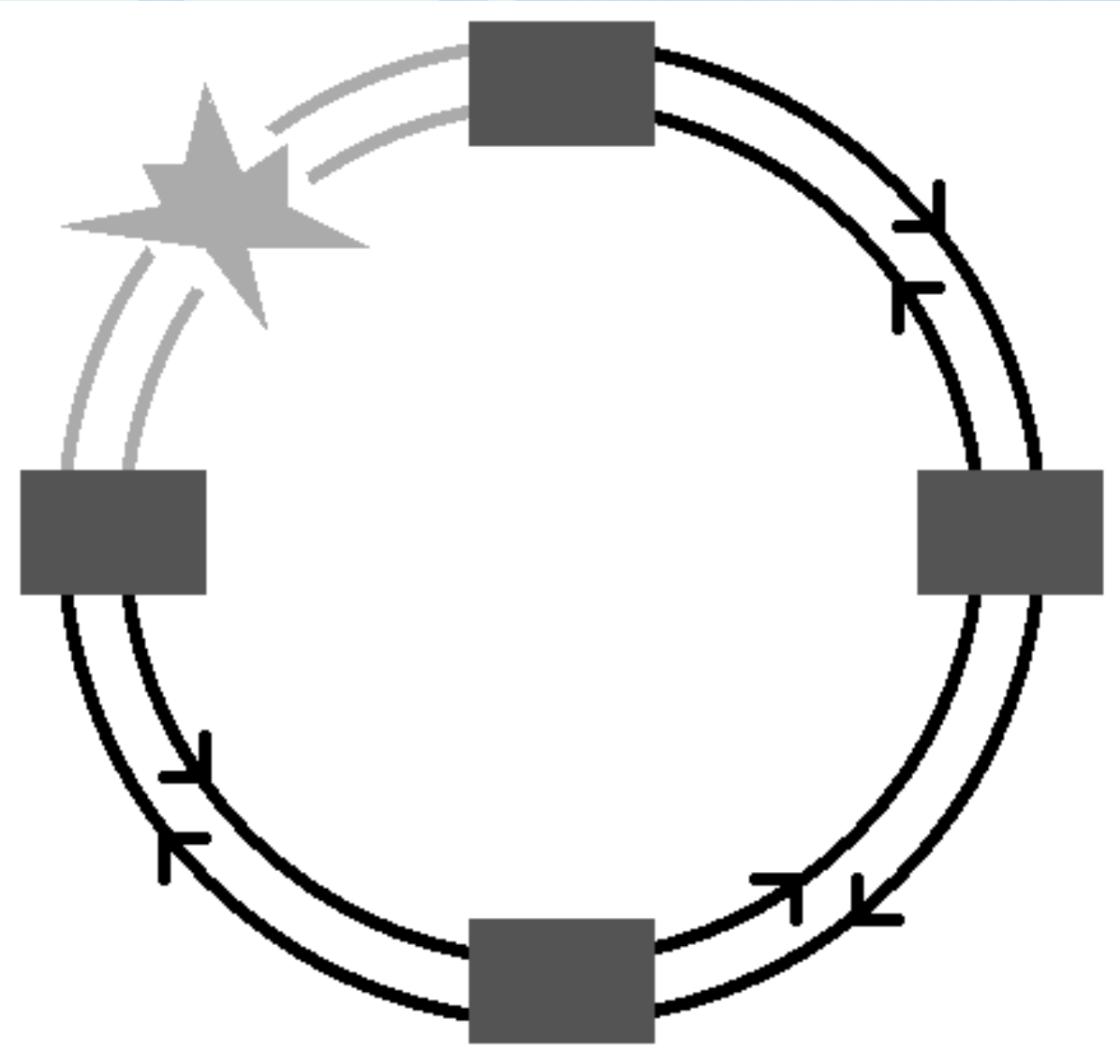
SONET/SDH

- Base is an "optical carrier" of 51.84 Mbps
- OC3/STM-1: 155 Mbps
- OC12/STM-4: 622 Mbps
- OC48/STM-16: 2488 Mbps
- OC192/STM-64: 9952 Mbps (129000 DS0s!)
- *c* is "concatenated": OC3c = 155, OC3 = 3 x 51

Protected SONET/SDH ring



"Autorepair" when fiber breaks



DWDM

- Until now:
  - Analog: Frequency Devision Multiplexing (FDM)
  - Digital: Time Devision Multiplexing (TDM)
- On fiber: (Dense) Wavelength Devision Multiplexing
- Different colors laser light through a single fiber
- 160 x 10 Gbps ??? (20 million DS0s)

We need fiber!

- Upto around 144 fibers in a single cable (1.5 billon DS0s)
- 8 of more ducts in a trench (12 billion DS0s)

The beginning: ARPANET

- ARPA: Defense Advanced Research Projects Agency
- Leased lines with 50 kbps modems (the size of a refrigirator), 4 locations in 1969
- Not for military use but for logging in to remote computers in different locations
- Packet switching and computer-to-computer rater than computer-to-terminal communication: revolutionary!

Packet switching

- Chop all forms of communication into small packets of about 1000 bytes
- Every packet contains destination address
- Paul Baran of the RAND Corporation "On Distributed Communications" memoranda
- Donald Watts Davies, National Physical Laboratory: "packet switching"
- AT&T: no way this is going to work

Growth!

- Enormous growth of the ARPANET in the 1970s
- Applications: at first only remote login (telnet) and file transfer (FTP), later also email
- Early 1980s: split Network Control Protocol into TCP (end-to-end) and IP (hop-by-hop)

ARPANET too successful...

- New network: National Science Foundation: 1544 kbps, between super computer locations
- In 1989 Federal Internet Exchanges to facilitate migration from ARPANET to NSFNET backbone
- NSFNET backbone Acceptable Use Policy: "no for-profit activities"
- Commercial Internet Exchange (CIX) and later MAE East were used by commercial networks to exchange traffic

1995: commercial backbones

- Congestion in NSFNET backbone, just like what happened to the ARPANET
- No longer government role to run backbone networks: commercial backbones + very high speed Backbone Network Service for research
- 4 Network Access Points interconnect backbones: MAE East (Washington), Sprint NAP (New Jersey), PacBell NAP (Palo Alto) and Ameritech NAP (Chicago)

Relationship voice/data

- Until around 1990: data "on top of" voice infrastructure, data bandwidth limited compared to voice
- Around 1995: data and voice equal: bandwidth is similar
- From around 2000: voice dwarfed by data: voice bandwidth limited compared to data
- Future: exit voice infrastructure, everything over IP?

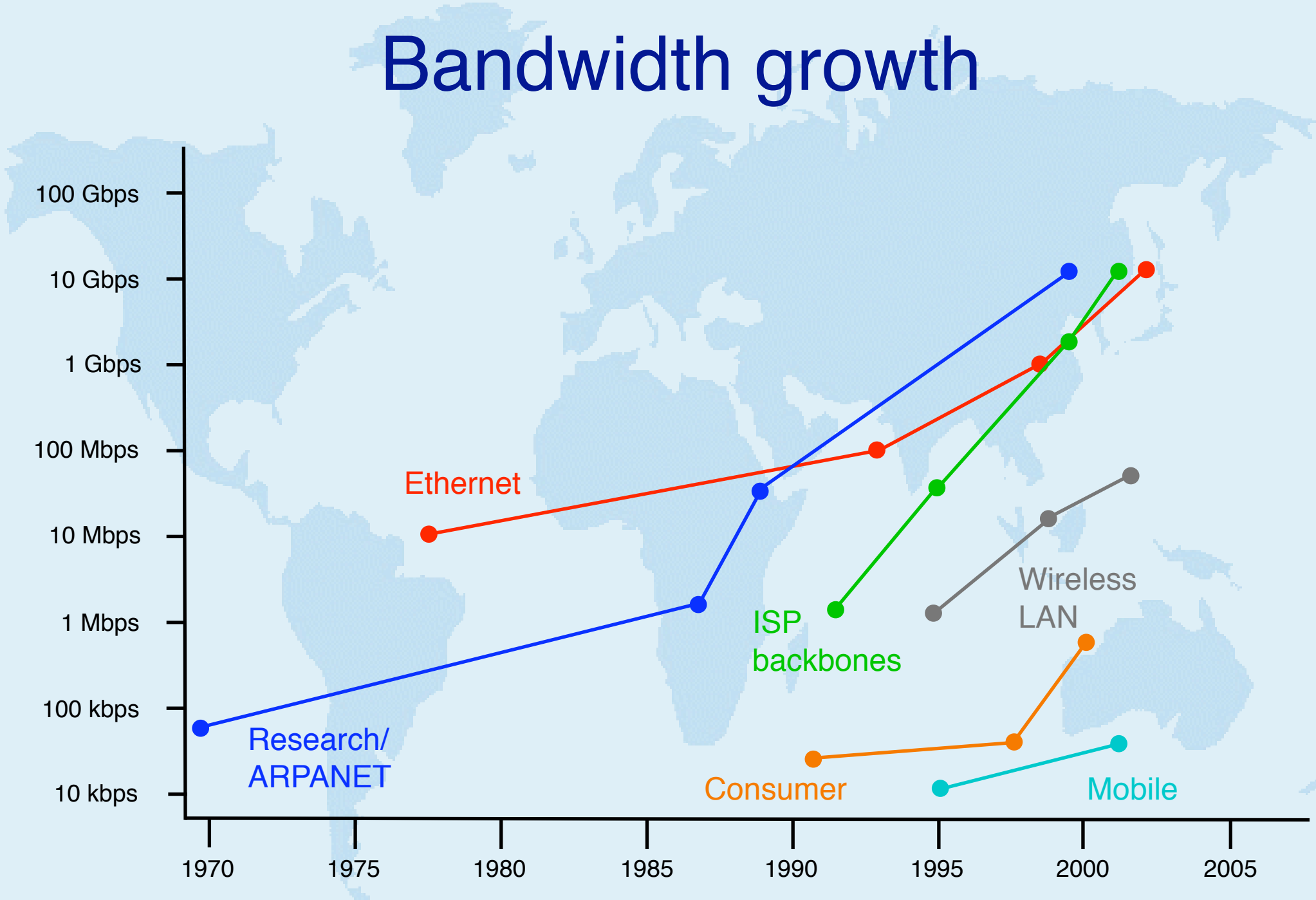
Fiber rings vs IP

- Of course IP can run over "protected" fiber rings
- But: two fibers in the ground while you only get to use one: waste of resources
- And: what if the SDH/SONET equipment breaks down? Still out in the cold despite fiber resiliency
- So: we'd rather have two less reliable, but independent links, routing protocols will take care of rerouting when there are failures

In the mean time in computerland

- Token Ring: 4 Mbps ring (mid 1970s)
- Ethernet: 10 Mbps shared medium (late 1970s)
- FDDI: 100 Mbps ring (mid 1980s)
- ATM: 155 Mbps point-to-point (early 1990s)
- Fast Ethernet: 100 Mbps shared (mid 1990s)
- Switched Ethernet: 10/100 star/p2p (late 1990s)
- Gigabit Ethernet: 1 Gbps star/p2p (late 1990s)

Bandwidth growth



In the mean time in homes

- New cabling infrastructure too expensive for residential use
- Phone network: point to point, no limitations on ISP choice, but bandwidth limited by phone infrastructure
- Cable TV networks: at first LAN paradigm but now virtual point to point network
- Asymmetric Digital Subscriber Line: boils down to leased line but again virtual p2p network

Wireless last mile?

- GSM/cell modems: too expensive, slow
- GPRS: a little faster, but often even more expensive!?!
- UMTS: isn't here yet, still relatively slow and expensive???
- Wifi: frequencies are free, everyone gets to play, getting faster all the time but limited range
- WLL (wireless local loop): where and when???

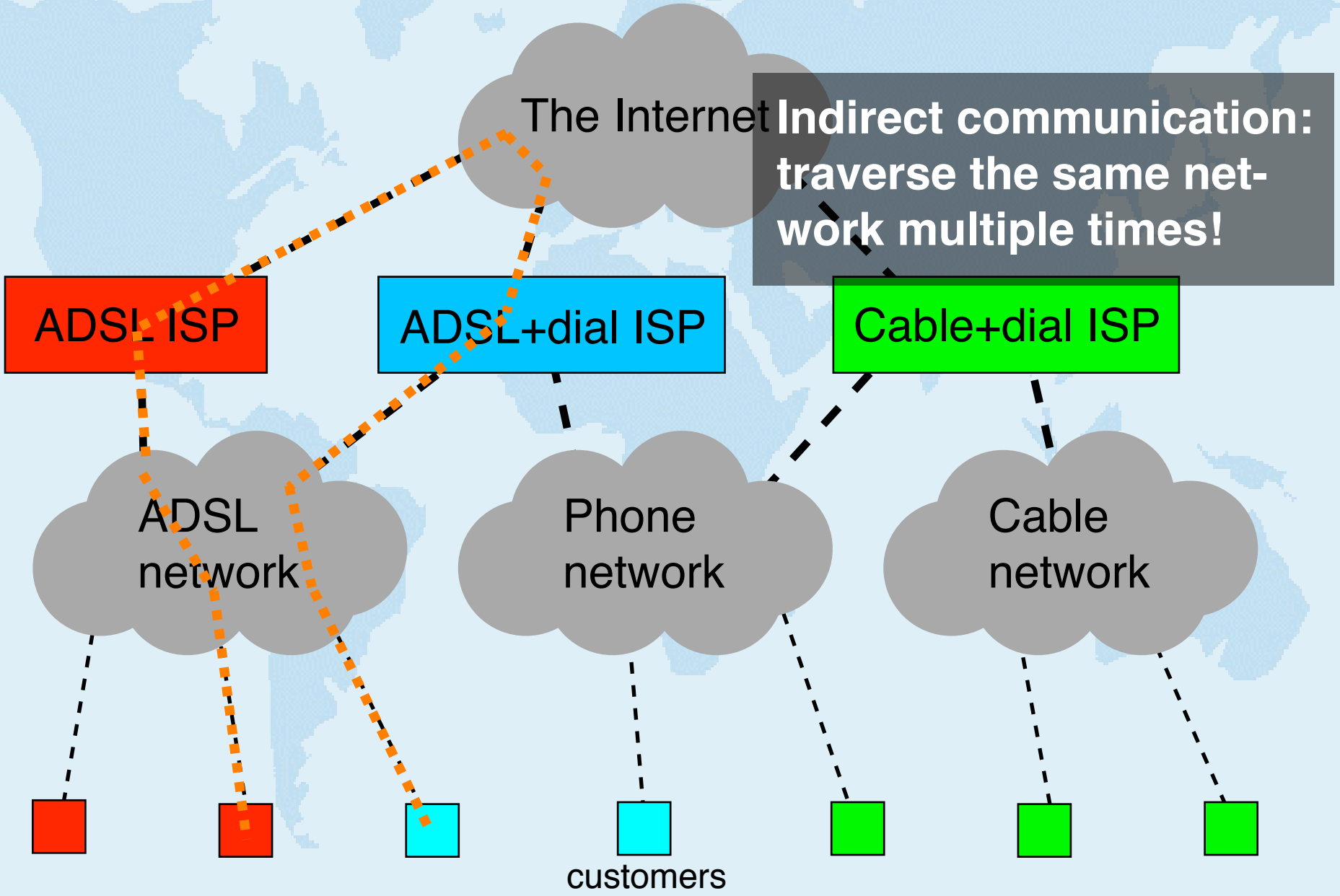
Fiber to the home

- Must be newly installed everywhere. Easy in high-rise apartment buildings?
- Only affordable if it also replaces cable TV and voice infrastructure
- Fiber not easy to handle
- Do we have enough capacity in backbone networks? Now at around 10 Gbps, that's equal to the aggregate capacity of 100 100 Mbps customers

Complexity

- ISP backbones now largely free of voice inheritance, running directly on top of fiber
- But end user infrastructure only gets more complex: IP over PPP over ATM or IP over PPP over Ethernet (over ATM...), IP over IP
- Complexity makes networks more expensive and less reliable, but often necessary for security and billing

Networks on top of networks



Internal routing protocols

- Within a clearly demarcated network (for instance, an enterprise or ISP network)
- Lets routers automatically find each other
- Each router tells others the IP address ranges it knows (because those are directly connected)
- Exchange information about "cost" of network links and determine cheapest/fastest path
- Choice of protocols: OSPF, RIP, EIGRP, IS-IS

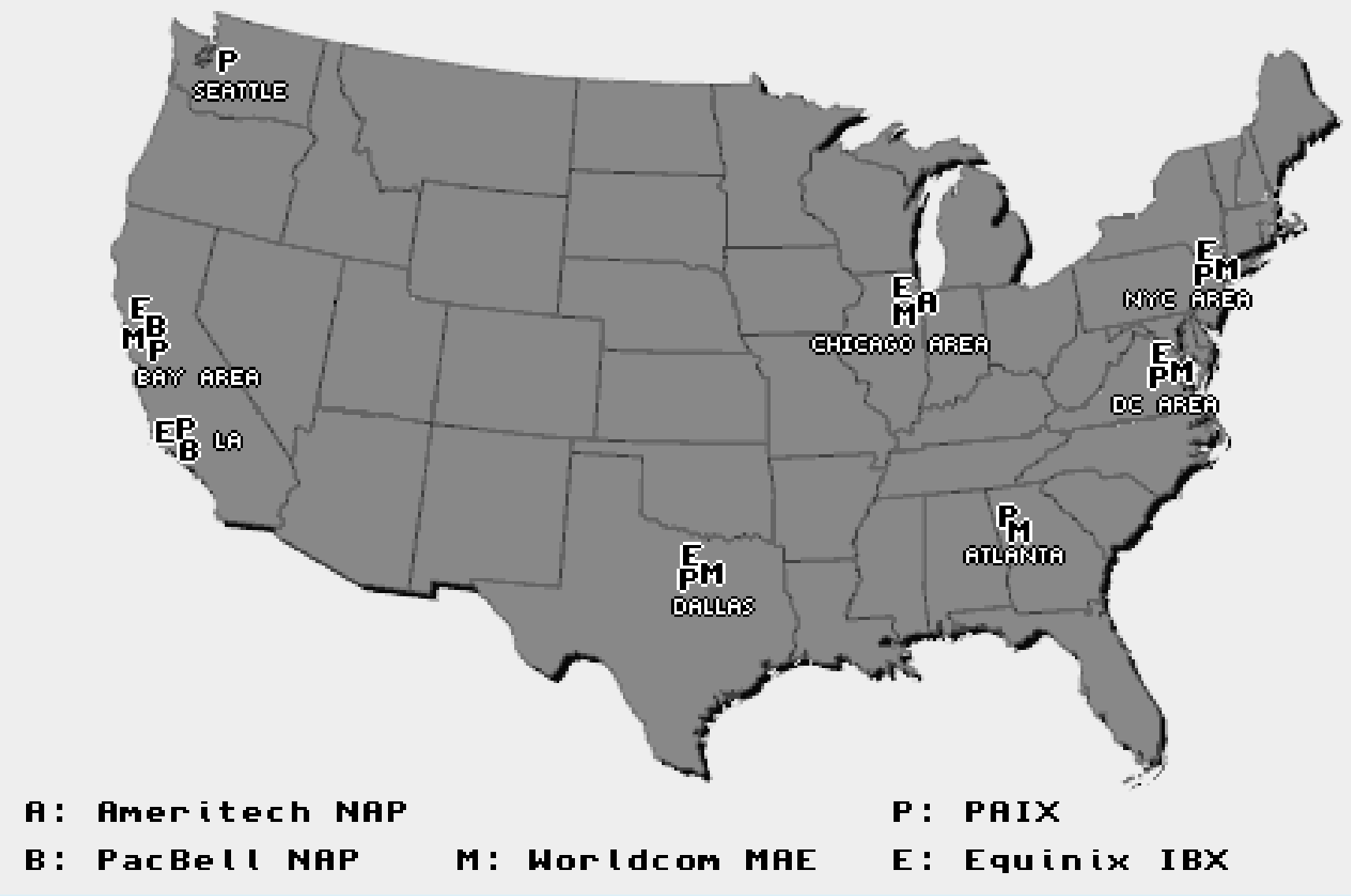
Between networks

- The Internet: network of networks
- So traffic must flow from one network to another
- ISPs interconnect through direct links (private interconnect) or internet exchanges

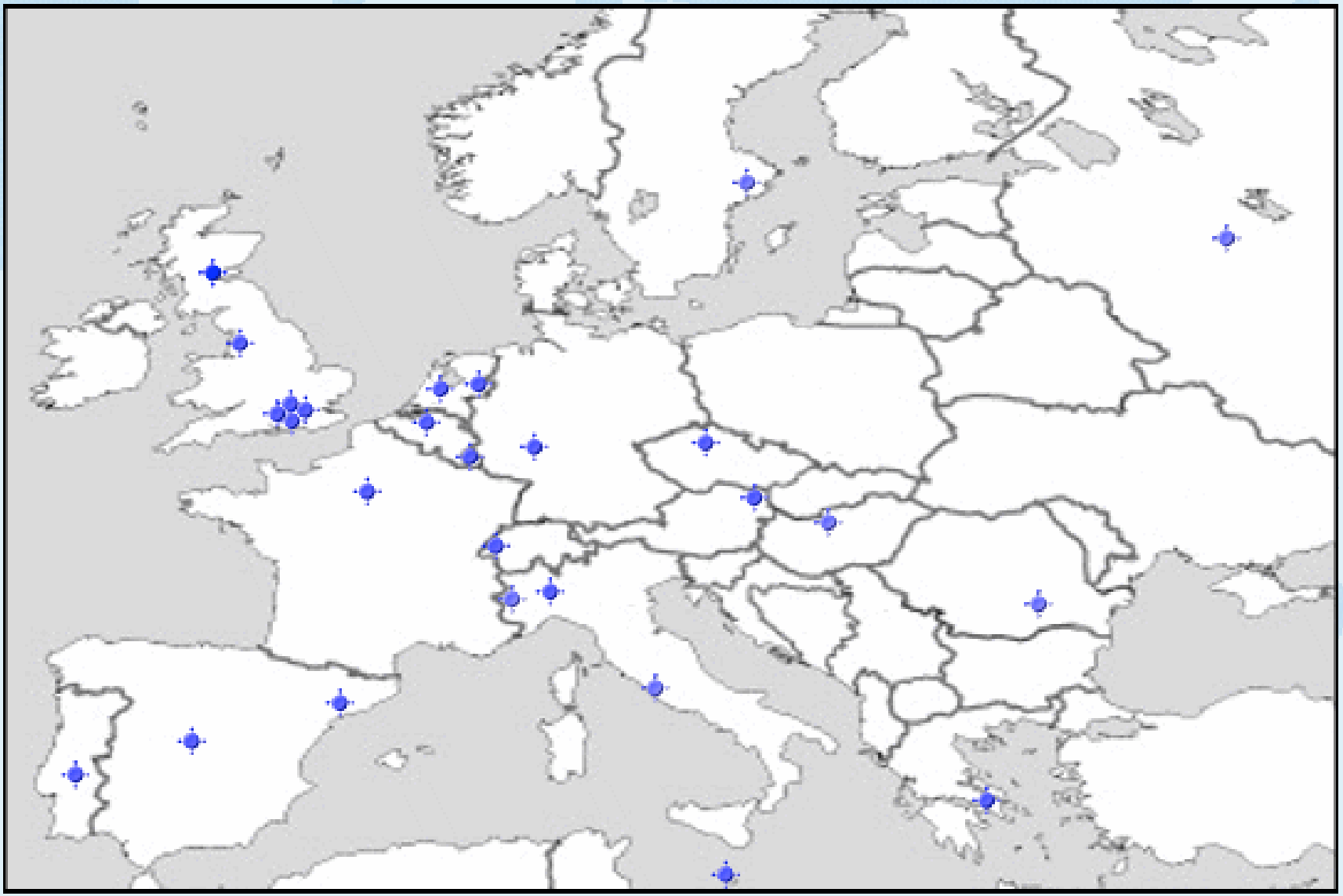
Internet exchanges

- Often a big Ethernet switch, sometimes ATM or something different
- In the US: many private interconnects, apart from that exchanges run by companies such as Worldcom, Equinix en PAIX and some small independent ones
- In Europe: pretty much every country has its own internet exchange, the big ones are independent

Interconnects US



Interconnects Europe



Routing between ISPs

- Internal routing protocols don't work here: too much information
- So: external routing protocols
- Another way of looking: not per router, but per network or organization, or rather "autonomous system"
- Only one external routing protocol in use: Border Gateway Protocol (BGP)

BGP's functions

- Pass on which IP addresses are use where
- Enforce "routing policy"
- Avoid routing loops
- Avoid broken links
- Additionally, when possible: choose shortest path

Which addresses go where

- No geographic relevance to IP address distribution
- Even then: geographically close locations may not be very close in network terms: two networks are often both be present in the same city but don't have an interconnect there
- So: explicitly list who is responsible for which address block so others know which ISP the packets should go to

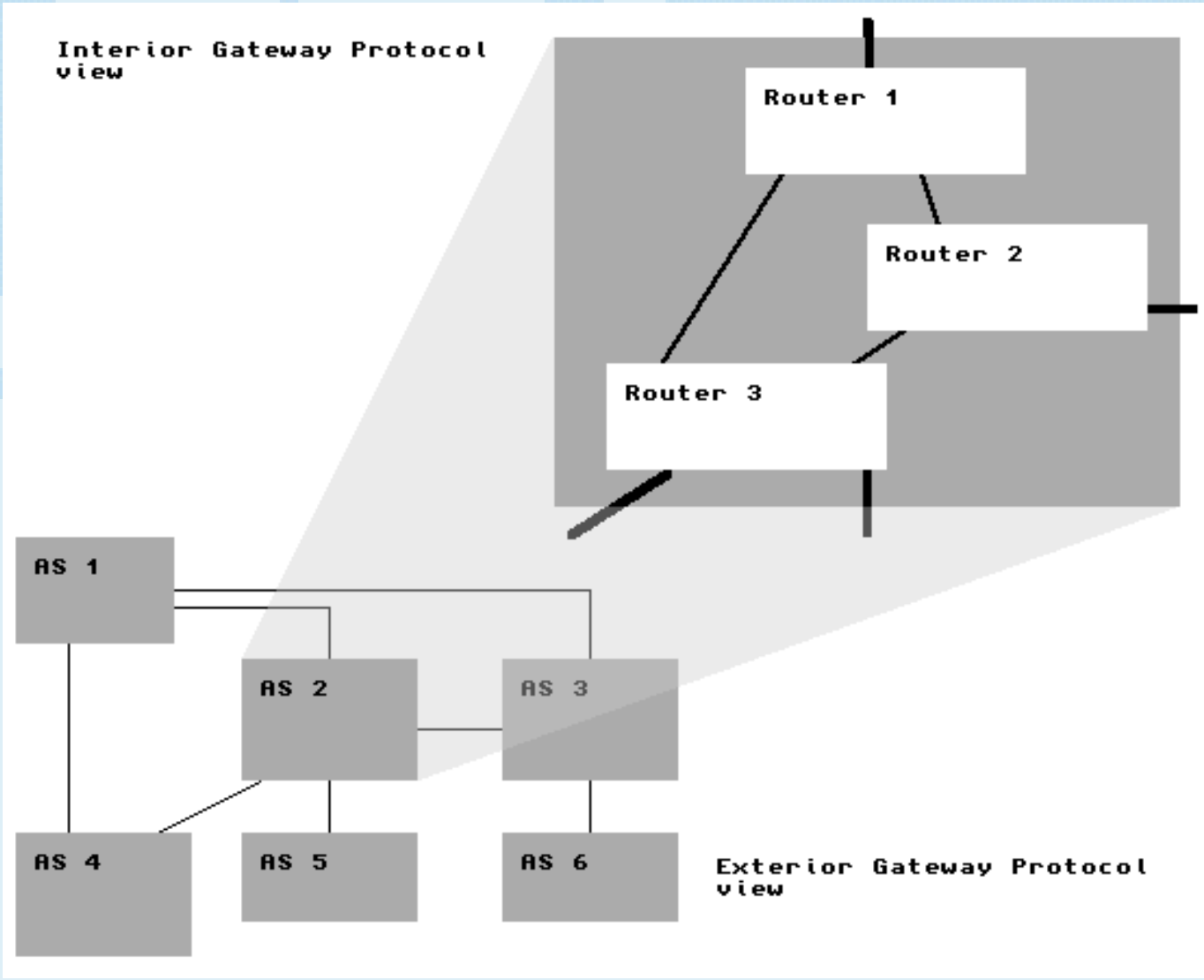
How BGP works

- "Border routers" have connections with border routers from neighboring ASes
- (And with all BGP routers within the local AS)
- Communication over TCP port 179
- Sessions are created manually

How BGP works (2)

- As soon as the connection is established each router sends a (more or less) complete copy of the "global routing table" to the neighbor
- Path through neighboring router better? Then use it yourself
- When this is done, only updates when something changes

Internal vs external view



The trees and the forrest

```
Tracing the route to www.isoc.nl (212.206.127.42)

  1 fa3-0-4-asd8ro2.enertel.nl (195.7.144.85) [AS 12394] 4 msec
  2 fa1-0-0-asd1ro6.enertel.nl (195.7.144.145) [AS 12394] 4 msec
  3 po0-0-0-asd10ro1.enertel.nl (195.7.154.14) [AS 12394] 4 msec
  4 adm-b2-pos2-1.telia.net (213.248.72.133) [AS 1299] 4 msec
  5 pos3-2.BR1.AMS3.ALTER.NET (146.188.64.113) [AS 702] 4 msec
  6 so-0-2-0.TR1.AMS2.ALTER.NET (146.188.3.213) [AS 702] 4 msec
  7 so-5-0-0.XR1.AMS6.ALTER.NET (146.188.8.77) [AS 702] 4 msec
  8 so-0-0-0.cr1.hag1.alter.net (212.136.176.110) [AS 702] 4 msec
  9 so-4-0-0.cr2.hag1.alter.net (212.136.176.146) [AS 702] 8 msec
 10 so-0-0-0.cr2.rtm1.alter.net (212.136.176.121) [AS 702] 8 msec
 11 412.atm10-0-0.gw4.rtm1.alter.net (212.136.177.146) [AS 702] 16 msec
 12 www.isoc.nl (212.206.127.42) [AS 702] 4 msec
```

The trees and the forrest (2)

```
Tracing the route to www.isoc.nl (212.206.127.42)

  1 fa3-0-4-asd8ro2.enertel.nl (195.7.144.85) [AS 12394] 4 msec
  2 fa1-0-0-asd1ro6.enertel.nl (195.7.144.145) [AS 12394] 4 msec
  3 po0-0-0-asd10ro1.enertel.nl (195.7.154.14) [AS 12394] 4 msec
  4 adm-b2-pos2-1.telia.net (213.248.72.133) [AS 1299] 4 msec
  5 pos3-2.BR1.AMS3.ALTER.NET (146.188.64.113) [AS 702] 4 msec
  6 so-0-2-0.TR1.AMS2.ALTER.NET (146.188.3.213) [AS 702] 4 msec
  7 so-5-0-0.XR1.AMS6.ALTER.NET (146.188.8.77) [AS 702] 4 msec
  8 so-0-0-0.cr1.hag1.alter.net (212.136.176.110) [AS 702] 4 msec
  9 so-4-0-0.cr2.hag1.alter.net (212.136.176.146) [AS 702] 8 msec
 10 so-0-0-0.cr2.rtm1.alter.net (212.136.176.121) [AS 702] 8 msec
 11 412.atm10-0-0.gw4.rtm1.alter.net (212.136.177.146) [AS 702] 16 msec
 12 www.isoc.nl (212.206.127.42) [AS 702] 4 msec
```

Policy

- Only send routes when allowed: don't provide services unless paid for
- Additionally: configured preference
- And: check whether what the neighbor sends is correct

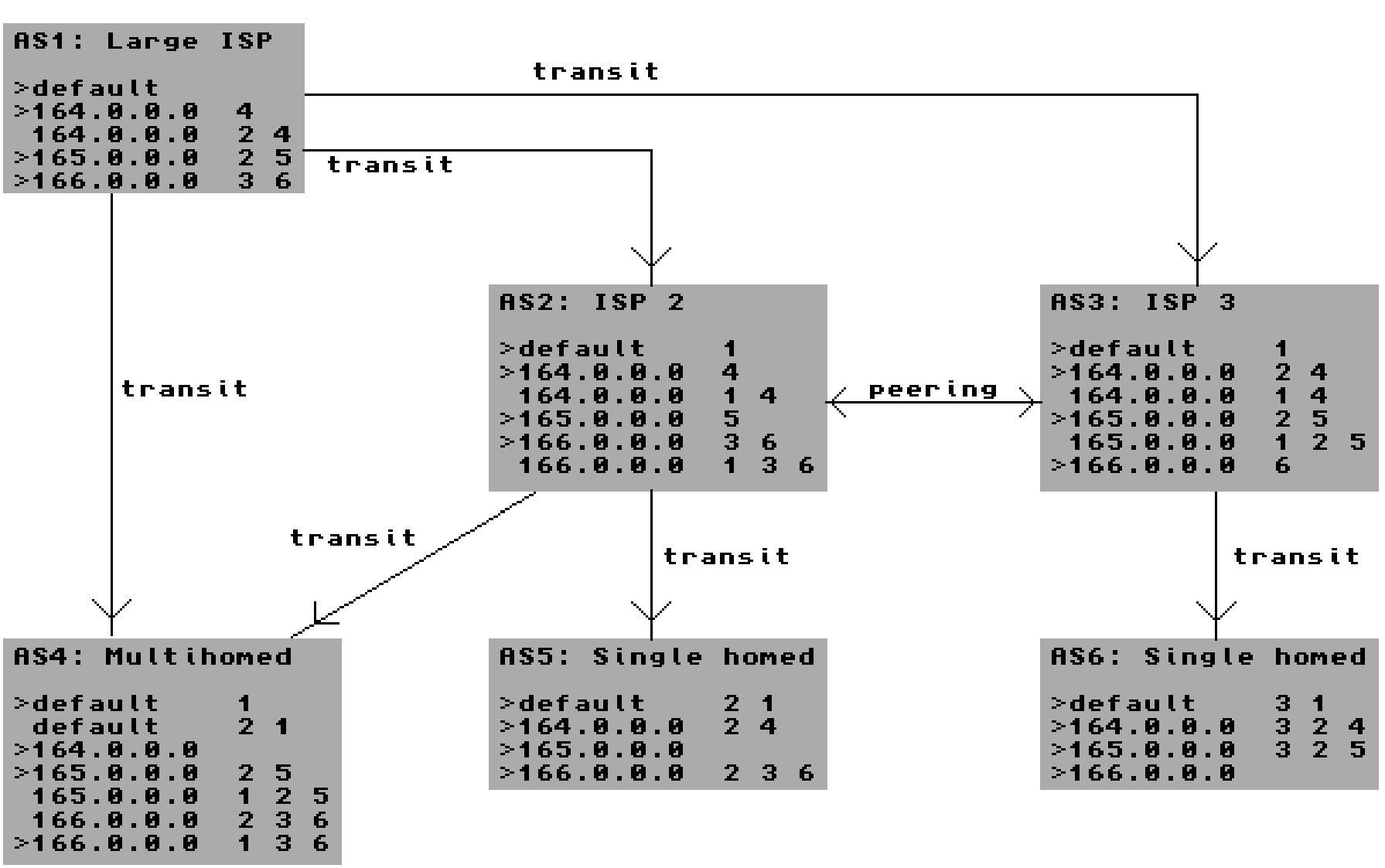
Transit

- Transit is provided to paying customers
- X provides transit to Y
- X routes packets from/to far away on behalf of Y
- So X announces to everyone that Y is reachable through them
- And X announces to Y that the whole world is reachable through them

Peering

- Exchange traffic without money changing hands (usually)
- X peers with Z
- X announces to Z that X's customers are reachable through X, but NOT the rest of the world
- Z does the same. Net effect: all traffic has X or a customer of X as its source and Z or a customer of Z as its destination (or the other way around)

Peering and transit



Path attributes

- Information routers exchange consists of a range of IP addresses and "path attributes"
- Address ranges take the shape of a prefix: 10.0.0.0/8, 192.168.0.0/16, 127.0.0.1/32
- Some common path attributes are:

  AS path, next hop, origin, Multi Exit Discriminator (or metric) and communities

AS path

- Mandatory
- First of all: detect and avoid routing loops
- Also important to enforce transit/peering filters
- And for filters that make sure customers don't accidently act as transit ISPs
- Shortest (allowed) path is preferred when comparing different routes for the same destination

Local preference

- Only exchanged within the local AS, but mandatory.
- Route with the highest local preference is used.
- Only when local pref is equal other attributes such as AS path are evaluated

Multi Exit Discriminator

- Optional
- Similar to "metric" in internal routing protocols
- Originally meant to differentiate between multiple similar routes over the same neighboring AS, but useful elsewhere too
- Isn't communicated to external ASes

Communities

- 32 bit value used to convey user-defined information
- Not in the original BGP specification! (So obviously optional)
- Some "well-known" communities
- Most take the shape of AS:nn (such as 701:120) where meaning depends on source AS
- Mostly used for special treatment of routes

In the wild...

```
   Network          Next Hop            Metric LocPrf Weight Path
*>i158.74.0.0    213.156.3.144           10    100        0 4589 1 i
*                 195.7.144.85             5               0 12394 3356 1 i
*> 158.94.0.0    195.7.144.85             5               0 12394 1299 786 i
*>i158.96.0.0    213.156.3.144           10    100        0 4589 3561 10840 i
*>i158.100.0.0   213.156.3.144           10    100        0 4589 3561 3908 i
*> 158.103.0.0   195.7.144.85             5               0 12394 1299 209 i
*> 213.17.3.0    195.7.144.85             5               0 12394 1299 9302 i
```

- "This is a route to all addresses that have 213.17.3 as their first 24 bits. The path to this network leads through ASes 12394, 1299 and 9302. The origin is IGP, the MED metric 5, there is no local preference present and send packets to the router with address 195.7.144.85."

Who uses BGP?

- Most ISPs, in order to tell other ISPs which address blocks they use
- End users?
  - Most don't, ISP handles routing for their addresses
  - End users with more than one ISP (who are said to be “multihomed”) do, since they can't depend on a single ISP to handle this

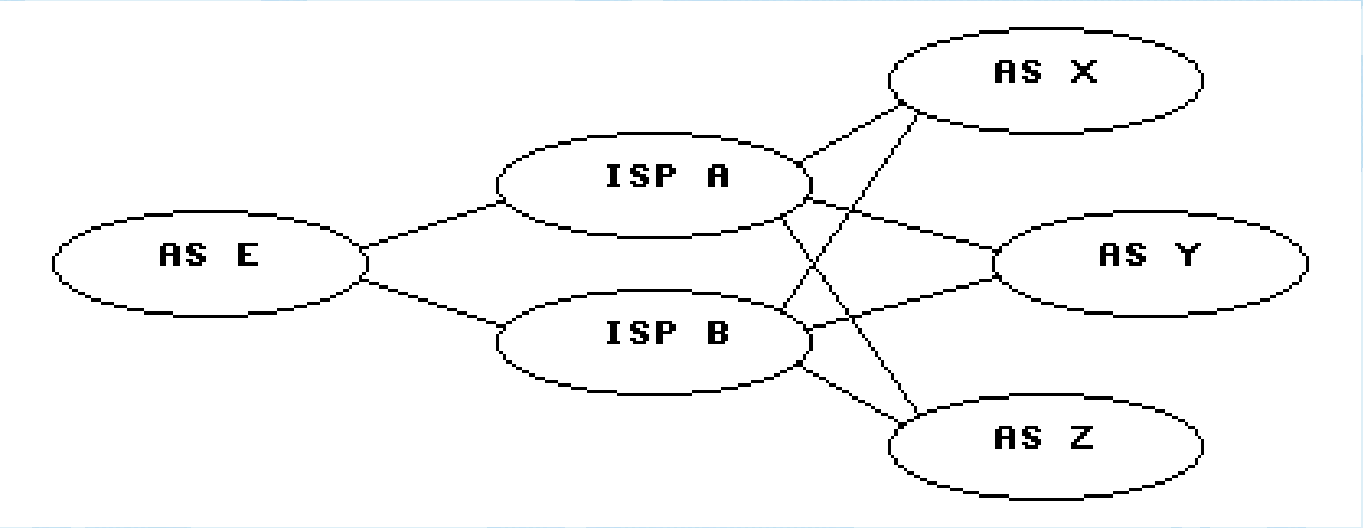
What BGP does for you

- Filter out certain routes:
  - On prefix (for instance, accept only previously agreed upon address blocks)
  - On AS path (for instance, only send routes with just the local AS and ASes from customers in the AS path to peers)
- Always prefer certain routes, like the ones learned over an internet exchange, by giving them a higher local preference

Balancing traffic

- When more than one external connection is available (without this BGP isn't very useful...) BGP automatically balances the traffic
- But not always in the most desirable way
- Influence this:
  - Make AS path longer for certain routes to make them less attractive
  - Set MED to prefer a route if AS paths are the same length

Sometimes too effective



- When two ISPs both peer with mostly the same networks paths will tend to be the same length
- In this case even small changes radically change traffic patterns

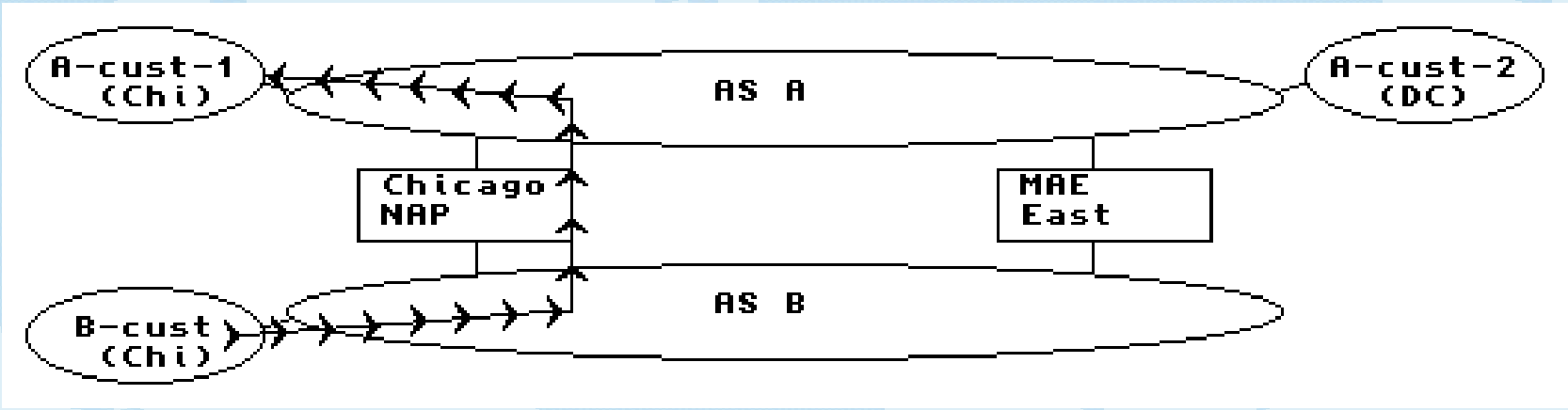
Communities in practice

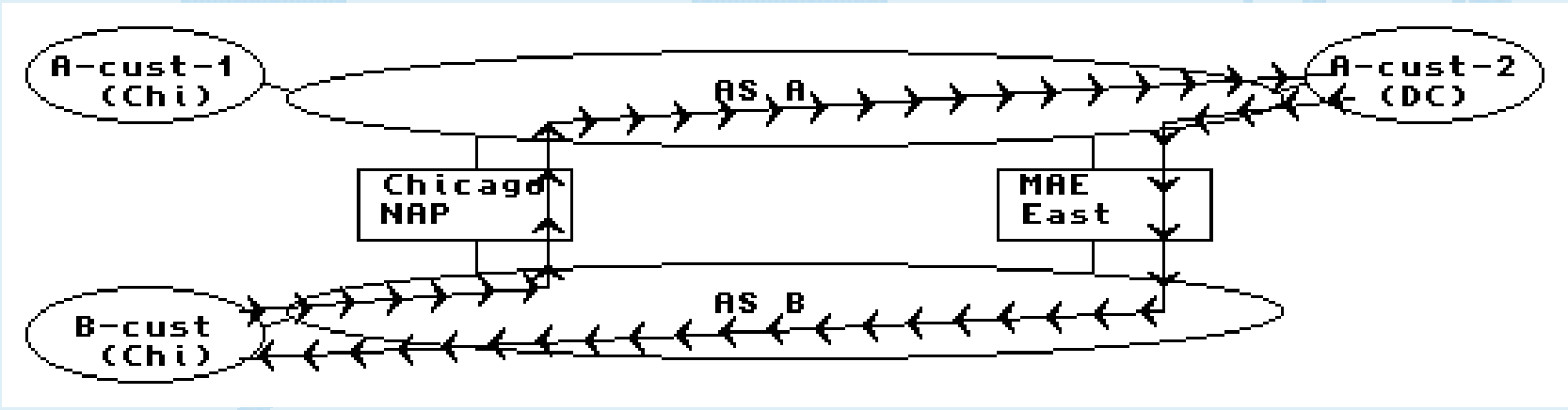
```
Aut-num:      AS702
as-name:      AS702
descr:        UUNET - Commercial IP service provider in Europe
import:       from AS72 194.98.169.195 at 194.98.169.196 accept AS72
import:       from AS109 213.53.49.50 at 213.53.49.49 accept AS109
[...]
export:       to AS72 194.98.169.195 at 194.98.169.196 announce ANY
export:       to AS109 213.53.49.50 at 213.53.49.49 announce ANY
[...]
remarks:      ------------------------------------------------------
remarks:      UUNET uses the following communities with its customers:
remarks:      702:80    Set Local Pref 80 within AS702
remarks:      702:120   Set Local Pref 120 within AS702
remarks:      702:20    Announce only to UUNET AS'es and UUNET customers
remarks:      702:30    Keep within Europe, don't announce to other UUNET AS's
remarks:      702:1     Prepend AS702 once at edges of UUNET to Peers
remarks:      702:2     Prepend AS702 twice at edges of UUNET to Peers
remarks:      702:3     Prepend AS702 thrice at edges of UUNET to Peers
remarks:      ------------------------------------------------------
```

How do ISPs use BGP

- Filter, filter, filter
- Well, maybe not all of them do this...
- Don't accept tiny address blocks (less than 256 addresses or /24): 120.000 prefixes is more than enough!
- "Hot potato" or "early exit" routing

Early exit: simple





Multiprotocol BGP

- Extension to BGP4 that makes it possible to distribute routing information for additional protocols
- Used for IP multicast: sending a single packet to more than one receiver
- And also IPv6, the new version of IP that allows for much more address space
- Some more esoteric uses such as VPNs

BGP security

- BGP is sensitive to attacks at the TCP, IP and link (usually ethernet) layers
- But in most cases not simple to exploit! MD5/ password offers protection, if hard to manage
- Future: Secure BGP (S-BGP) and/or secure origin BGP (soBGP)
- Heavy encryption too much? Most problems are due to human error anyway

Risks with current situation

- Filtering makes adapting to change harder
- Too much happens in too few places, see New York on 9/11
- Additional links often don't work as well as expected
- 120,000 routes is *a lot*, little room for error
- All those bankruptcies don't really help

That's it!

Iljitsch van Beijnum

iljitsch@bgpexpert.com
http://www.bgpexpert.com/

:-)