BGP

en de

fysieke infrastructuur

van het

internet

Iljitsch van Beijnum

Groningen/Amsterdam 23/24 april 2003

In telefonieland

- 1903: al 3,2 miljoen telefoonabonnees in de VS! Afstand beperkt tot ruim 2000 km
- 1915: buizenversterkers, nu ook echt lange afstanden over telefoondraden aan palen
- Jaren '50: lange afstand door multiplexing over microgolfradio en coax (analoog!)
- Vanaf jaren '60: transistors en digitaal

Digitale telefonie (of data)

- DS0 of "B kanaal": 64 kbps, 1 gesprek 8 kHz sampling rate en 8 bits: geen CD kwaliteit!
- T1 of DS1: 1544 kbps, 24 DS0s (meestal 1 "D kanaal" voor besturing)
- T2: huh? (6312 kbps)
- T3of DS3: 44736 kbps, 28 T1s of 672 DS0s

Europese standaarden

- DS0: 64 kbps, geen verschil
- E1: 2048 kbps, 31 DS0s of 30B+D
- E2: huh? (8448 kbps)
- E3: 34368 kbps, 16 E1s, 480/496 DS0s
- E4: 139264 kbps, 4 E3s

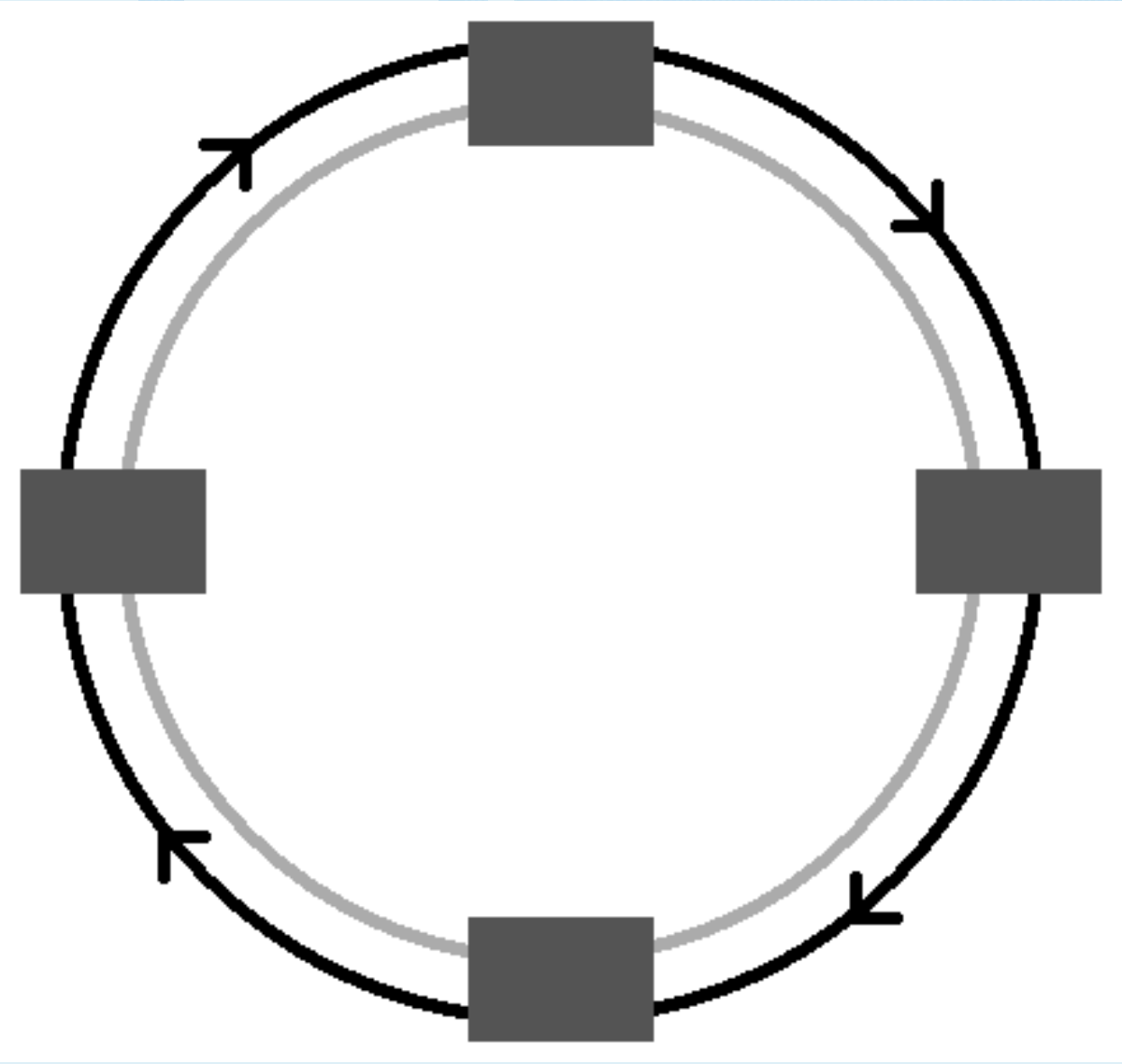
Kabels

- Telefoonlijn: unshielded twisted pair (UTP) koperdraad (analoog)
- T1/E1: afgeschermermde koperdraad (coax) of UTP
- T3/E3: coax
- Hogere snelheden: SONET/SDH 155, 622, 2488 of 9952 Mbps over glasvezel

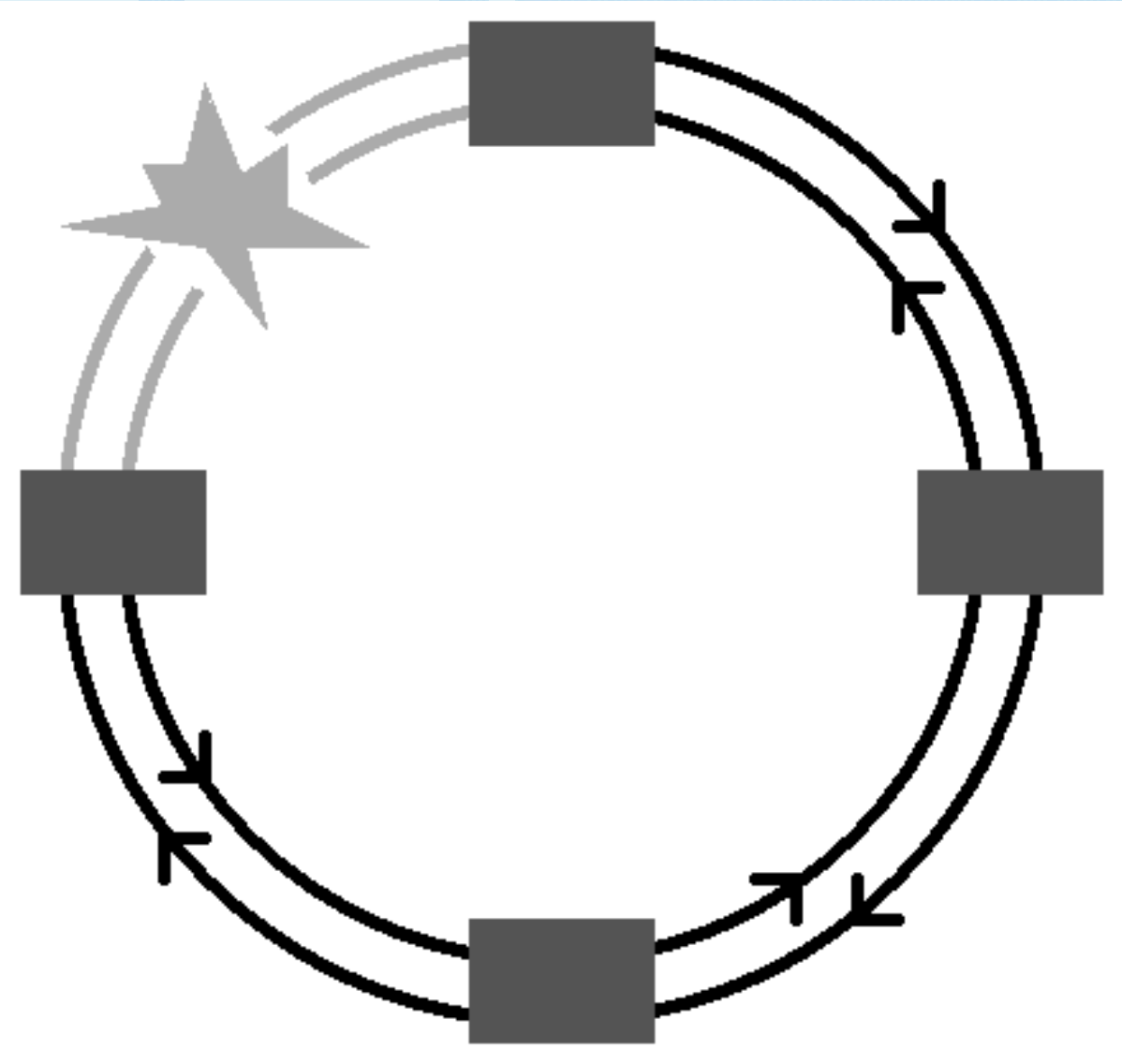
SONET/SDH

- Basis is een "optical carrier" van 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

SONET/SDH ring



"Autorepair" bij fiberbreuk



DWDM

- Tot nu toe: Time Devision Multiplexing (TDM)
- (Dense) Wavelength Devision Multiplexing
- Verschillende kleurtjes laserlicht door één fiber
- 160 x 10 Gbps ??? (20 miljoen DS0s)

We hebben glas nodig!

- Zo'n 144 fibers in een glasvezelkabel (1,5 miljard DS0s)
- 8 of meer buizen per geul (12 miljard DS0s)

Het begin: ARPANET

- ARPA: Defense Advanced Research Agency
- 4 locaties in 1969
- 50 kbps vaste verbindingen
- Geen militair doel maar gebruik op afstand van computers op verschillende locaties
- Packet switching en computer-naar-computer in plaast van computer-naar-terminal: revolutionair!

Packet switching

- Hak iedere vorm van communicatie in kleine pakketjes van zo'n 1000 bytes
- Bestemming zit in ieder pakket
- Paul Baran van de RAND Corporation "On Distributed Communications" memoranda
- Donald Watts Davies, National Physical Laboratory: "packet switching"
- AT&T: no way!

Groei!

- Onstuimige groei van het ARPANET in de jaren '70
- Applicaties: eerst alleen login (telnet) en file transfer (FTP), later ook mail
- Begin jaren '80: splitsing Network Control Protocol in TCP (end-to-end) en IP (hop-by-hop)

ARPANET te succesvol...

- Nieuw netwerk: National Science Foundation: 1544 kbps, supercomputer-locaties
- In 1989 Federal Internet Exchanges voor overgang ARPANET naar NSFNET backone
- NSFNET backbone Acceptable Use Policy: "no for-profit activities"
- Commercial Internet Exchange en later MAE East waar commerciele netwerken aansluiten

1995: de commercie

- Congestie in NSFNET backbone, net als eerder ARPANET
- Geen overheidstaak maar commerciele backbones + very high speed Backbone Network Service
- 4 Network Access Points om backbones te verbinden: MAE East (Washington), Sprint NAP (New Jersey), PacBell NAP (Palo Alto) en Ameritech NAP (Chicago)

Verhouding telefonie/data

- Tot ca. 1990: data "bovenop" het telefoonnetwerk, datasnelheden beperkt in telefonietermen
- Rond 1995: data "naast" het telefoonnetwerk, snelheden data - telefonie vergelijkbaar
- Vanaf ca. 2000: "achter" het telefoonnetwerk, telefoniesnelheden beperkt in datatermen
- Toekomst: exit telefoonnetwerk en alles over IP?

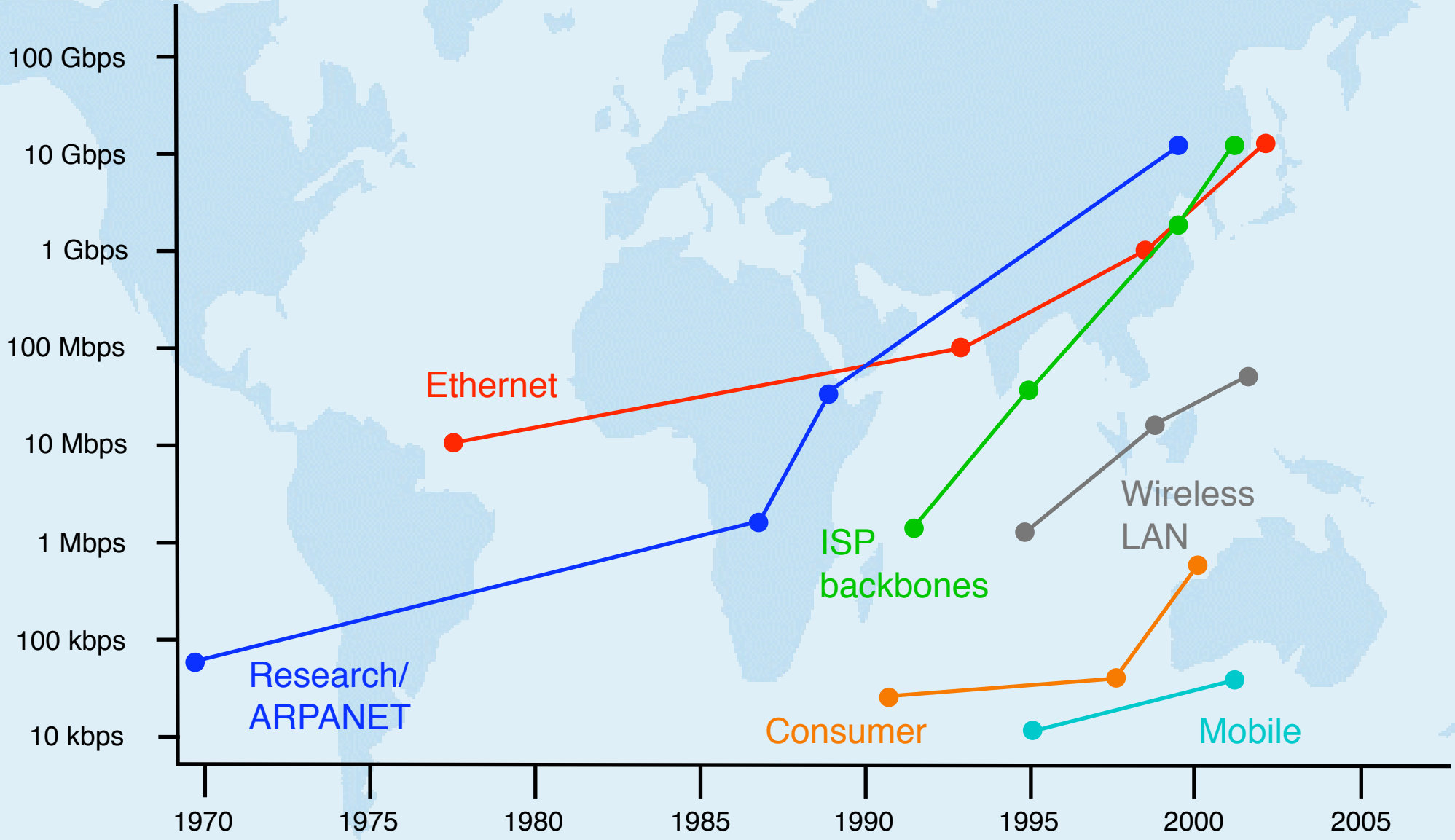
Fiberringen vs IP

- Uiteraard kunnen IP-netwerken over beschermde fiberringen draaien
- Maar: twee vezels in de grond terwijl je er maar een gebruikt, zonde!
- Wat als de SDH apparatuur kapot gaat?
- Dus liever twee onbetrouwbare verbindingen dan een betrouwbare, routing protocollen zorgen voor de rest

Intussen in computerland...

- Token Ring: 4 Mbps ring (midden '70)
- Ethernet: 10 Mbps bus (eind '70)
- FDDI: 100 Mbps ring (midden '80)
- ATM: oa 155 Mbps point-to-point (begin '90)
- Fast Ethernet: 100 Mbps ster/p2p/bus (midden '90)
- Gigabit Ethernet: 1000 Mbps ster/p2p (eind '90)

Groei bandbreedte



Intussen bij de mensen thuis...

- Voor thuisgebruik kabels aanleggen te duur
- Telefonienetwerk: inbellen bij veel verschillende ISPs, maar beperkt door bandbreedte telefoonnet
- Kabeltelevisienet: sneller, maar bandbreedte moet gedeeld worden. In het begin LAN-axioma, later virtueel netwerk
- Asymmetric Digital Subscriber Line: in feite huurlijn, maar weer virtueel netwerk

Last mile draadloos?

- GSM: duur, langzaam
- GPRS: iets minder langzaam, maar soms nog duurder!?!
- UMTS: is er nog niet, duur, relatief langzaam???
- wifi: geen dure frequenties, iedereen kan hobbyen, steeds sneller, zeer beperkt bereik
- Wireless local loop: niet mobiel, wel draadloos, waar blijft het?

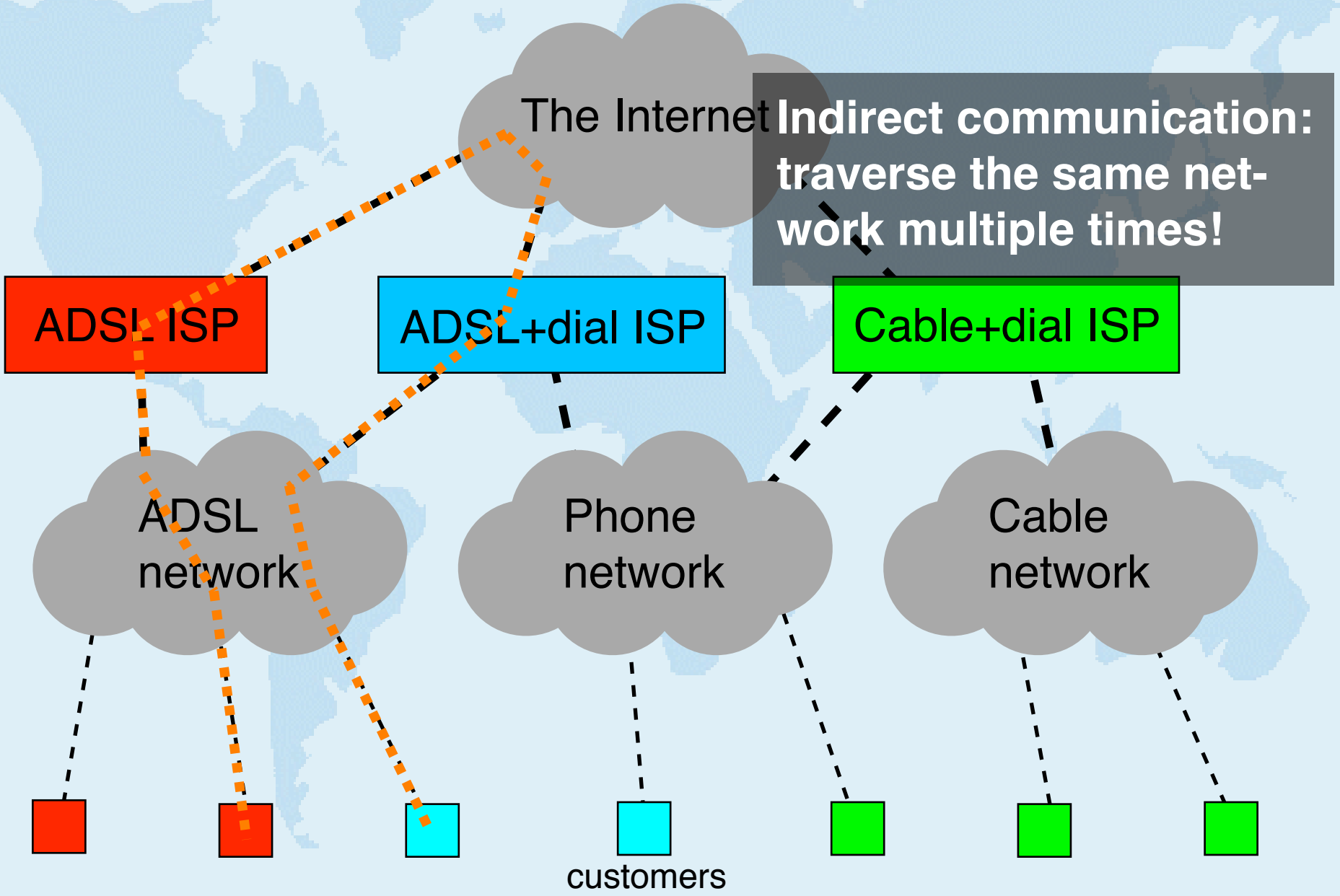
Fiber to the home

- Moet nieuw aangelegd worden, makkelijk in nieuwbouw, flats???
- Alleen betaalbaar als ook voor RTV en telefonie
- Glasvezel lastig te verwerken
- Hebben we wel zoveel capaciteit in de backbone?

Complexiteit

- Backbones hebben zich grotendeels vrijgemaakt van onderliggende netwerken en draaien rechtstreeks op fiber
- Bij eindgebruikers juist steeds meer complexe onderliggende zaken: PPP over Ethernet, Point-to-Point Tunneling Protocol, Virtual Private Networks
- Complexiteit maakt netwerken duur en onbetrouwbaar, maar vaak nodig voor beveiliging en billing

Networks on top of networks



Interne routingprotocollen

- Binnen een goed afgebakend netwerk (bv een ISP)
- Laat routers elkaar detecteren
- Vertel andere routers welke IP adressen waar aangesloten zijn
- Wissel gegevens over de "kosten" van verbindingen uit
- Bepaal de route met het "goedkoopste" pad

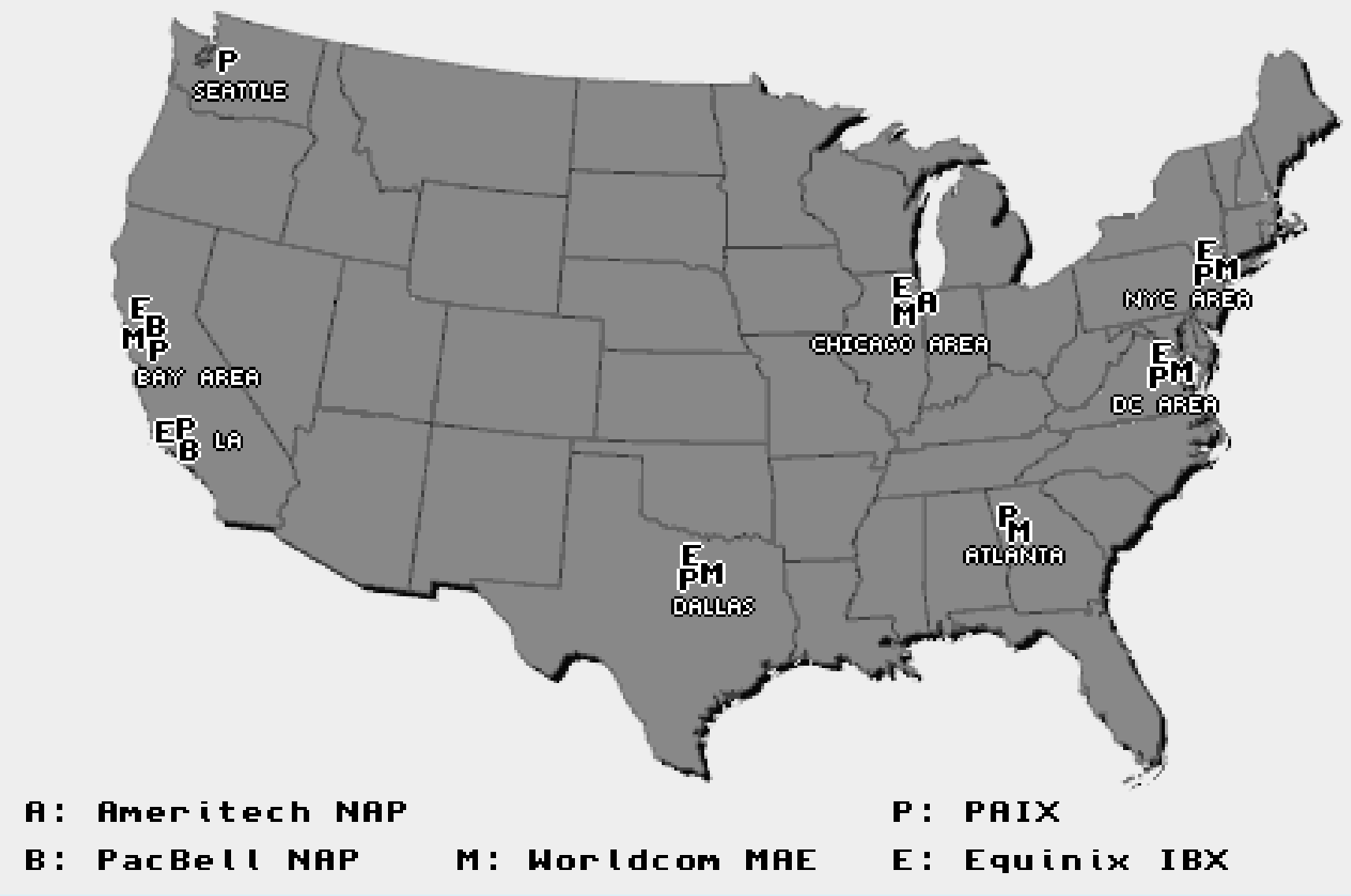
Tussen netwerken

- Het internet: netwerk van netwerken
- Dus overspringen van het ene netwerk naar het andere
- ISPs koppelen aan elkaar via directe verbindingen (private interconnect) of internet exchanges

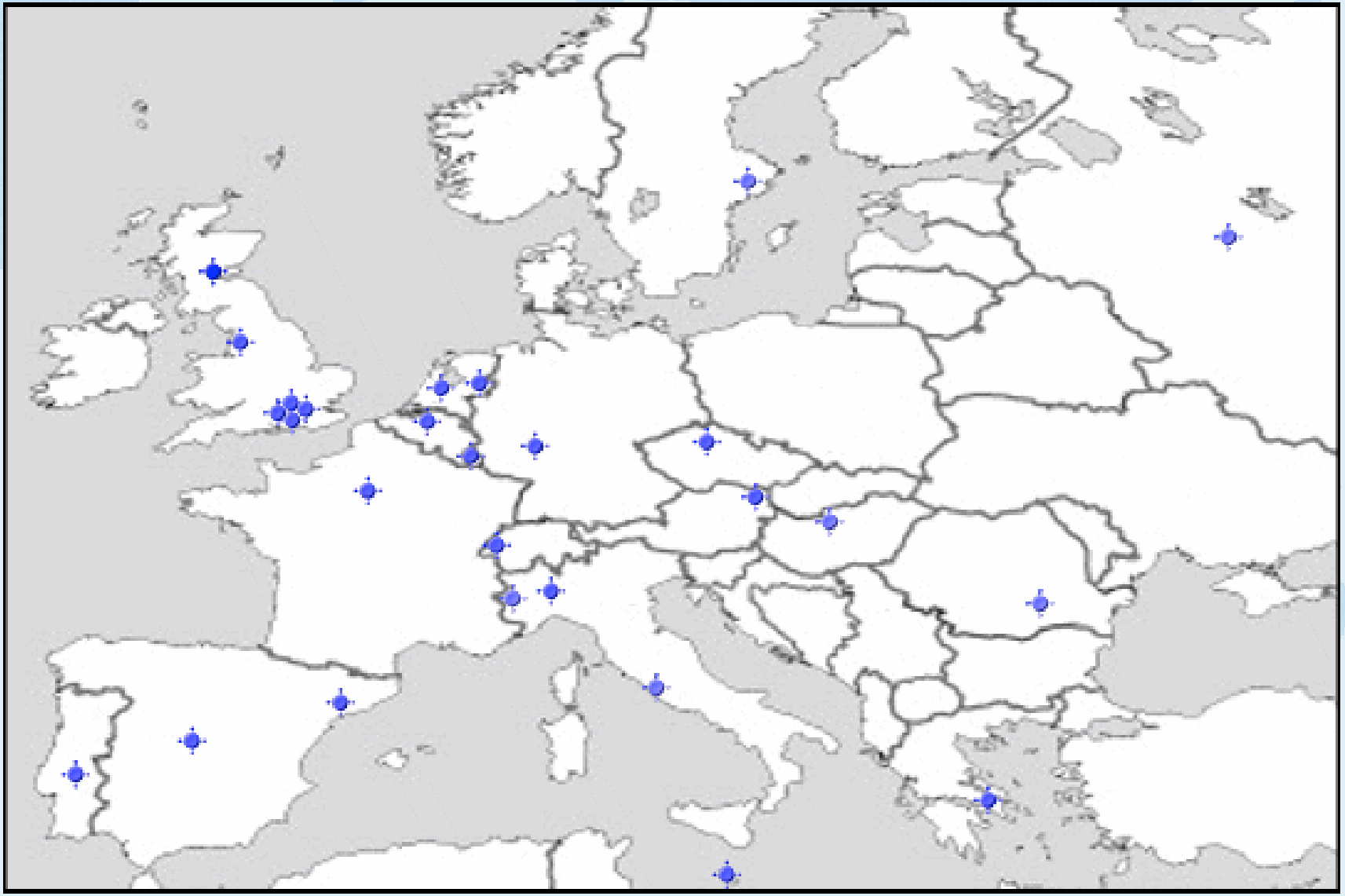
Internet Exchanges

- Meestal grote ethernet switch, soms ATM of anders
- In Amerika: veel private interconnects, kleine exchanges en commerciele bedrijven zoals Worldcom, Equinix en PAIX
- In Europa: ieder land heeft wel een eigen internet exchange, en enkele grote, meestal onafhankelijk

Interconnects VS



Interconnects Europa



Routering tussen ISPs

- Interne routingprotocollen werken hier niet: teveel informatie
- Dus: externe routing protocollen
- Andere manier van kijken: niet per router, maar per netwerk of organisatie ofwel "autonomous system"
- Maar één protocol: Border Gateway Protocol

Functies BGP

- Doorgeven welke IP adressen waar gebruikt worden
- "Policy" handhaven
- Routing loops voorkomen
- Kapotte verbindingen omzeilen
- En eventueel nog: kortste pad kiezen

Welke adressen waar

- Geen geografische relevantie aan IP adressen
- Zelfs dan: weinig geografische relevantie netwerktopologie
- Dus: expliciet aangeven wie welk adresblok "onder zich" heeft om de pakketten naar de goede ISP te kunnen sturen

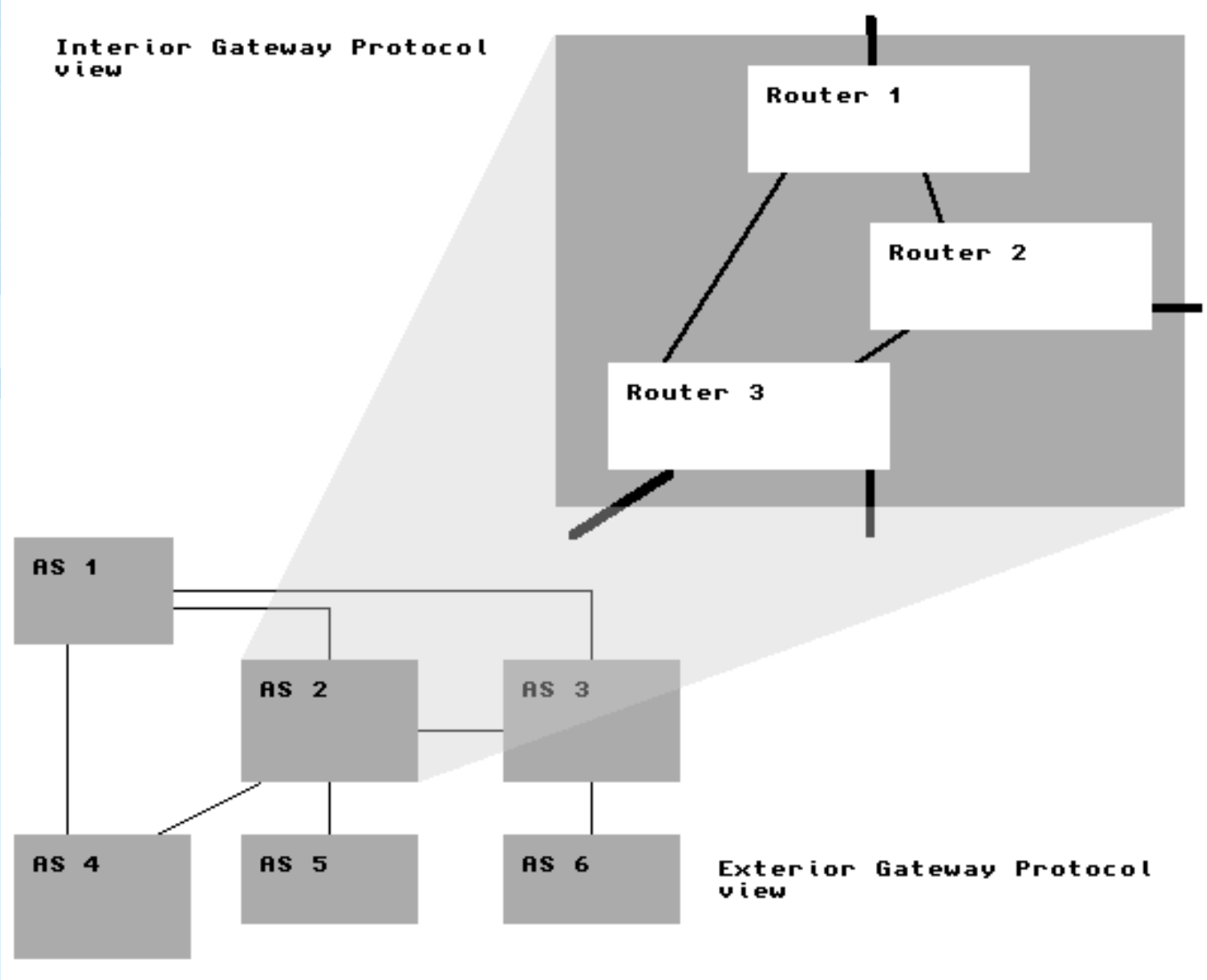
Hoe BGP werkt

- "Border routers" onderhouden een verbinding met de border routers van naburige ASen
- (En met alle andere border routers binnen het eigen AS)
- Communicatie over TCP poort 179
- Sessies worden handmatig aangemaakt

Hoe BGP werkt (2)

- Zodra de verbinding opkomt stuurt iedere router een (min of meer volledige) kopie van de "global routing table" naar de buur
- Via buur-router beter? Gebruik dit pad dan zelf
- Zodra dit gebeurt is alleen updates als er wat verandert

Intern vs extern



De bomen en het bos

```
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
```

De bomen en het bos (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

- Maar wel alleen als het mag: geen service leveren aan iemand die niet betaalt
- Daarnaast: ingestelde voorkeur/nakeur
- En: wel zeker weten dat wat buren doen klopt

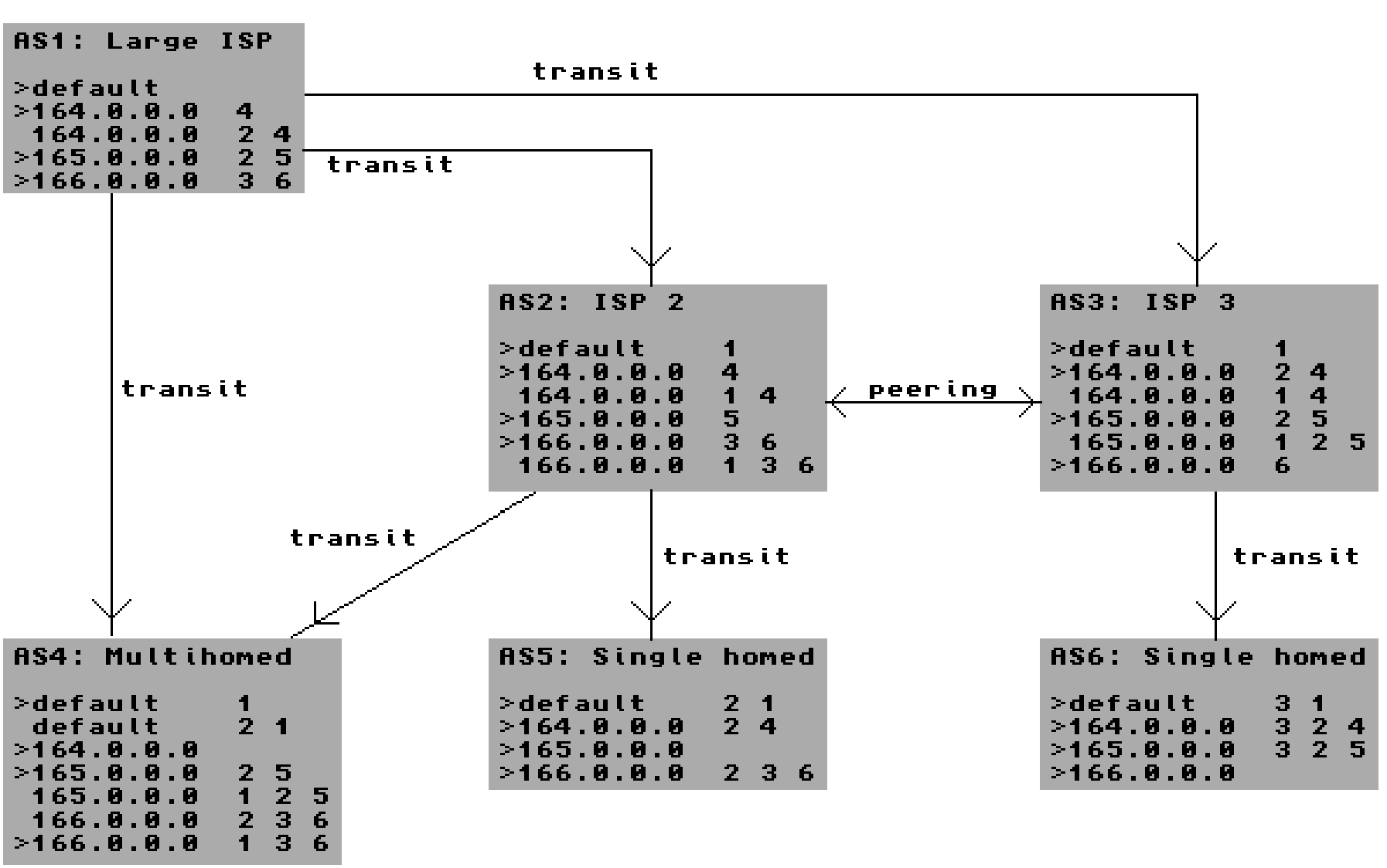
Transit

- Transit lever je aan betalende klanten
- X levert transit aan Y
- X routeert pakketten van/naar ver weg voor Y
- Dus X geeft aan iedereen door dat Y via hem bereikbaar is
- En X geeft aan Y door dat de hele wereld via hem bereikbaar is

Peering

- Uitwisselen van verkeer zonder dat er geld aan te pas komt (meestal)
- X peert met Z
- X geeft aan Z aan dat klanten via hem bereikbaar zijn, maar NIET de rest van de wereld
- Z doet hetzelfde. Netto effect: al het verkeer heeft X of een klant van X als bron en Z of een klant van Z als bestemming (of omgekeerd)

Peering en transit



Path attributes

- Informatie die routers uitwisselen bestaat uit een reeks IP adressen en "path attributes"
- Adressen in de vorm van een prefix: 10.0.0.0/8, 192.168.0.0/16, 127.0.0.1/32
- Path attributes zijn onder andere:
  - AS pad
  - Next hop
  - Origin

AS pad

- Allereerst: routing loops te voorkomen
- Daarnaast belangrijk in filters om transit/peering af te dwingen en te voorkomen dat klanten per ongeluk transit gaan leveren
- Kortste pad gaat voor in route selectie

Local preference

- Wordt alleen binnen het AS doorgegeven, maar wel verplicht
- Route met de hoogste local preference wordt gebruikt. Alleen als local pref gelijk is wordt naar AS pad en andere zaken gekeken

Multi Exit Discriminator

- Vergelijkbaar met "metric" in interne routingprotocollen
- Oorspronkelijk alleen om doorslag te geven bij meerdere routes van hetzelfde AS, maar meestal ook wel breder inzetbaar
- Wordt niet verder dan het volgende AS doorgegeven

Community's

- Niet in de originele BGP specificatie!
- Maar enkele "well-known" community's
- Over het algemeen: AS:nn
- Betekenis hangt van de zender/ontvanger af, meestal gebruikt voor speciale behandeling

In het echt...

```
   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
```

- "Hier is een route naar alle adressen waarvan de eerste 24 bits 213.17.3.0 zijn. Het pad ernaartoe bevat de ASen 12394, 1299 en 9302. De MED metric is 5, er is geen local preference aanwezig en stuur de pakketten naar 195.7.144.85."

BGP voor wie?

- ISPs, in elk geval de grotere, om adressen aan andere ISPs door te geven
- Eindgebruikers?
- De meeste niet, liften mee met ISP
- Meerdere ISPs (“multihoming”) wel, kan immers niet afhankelijk zijn van de ISP die adressen de wereld instuurt

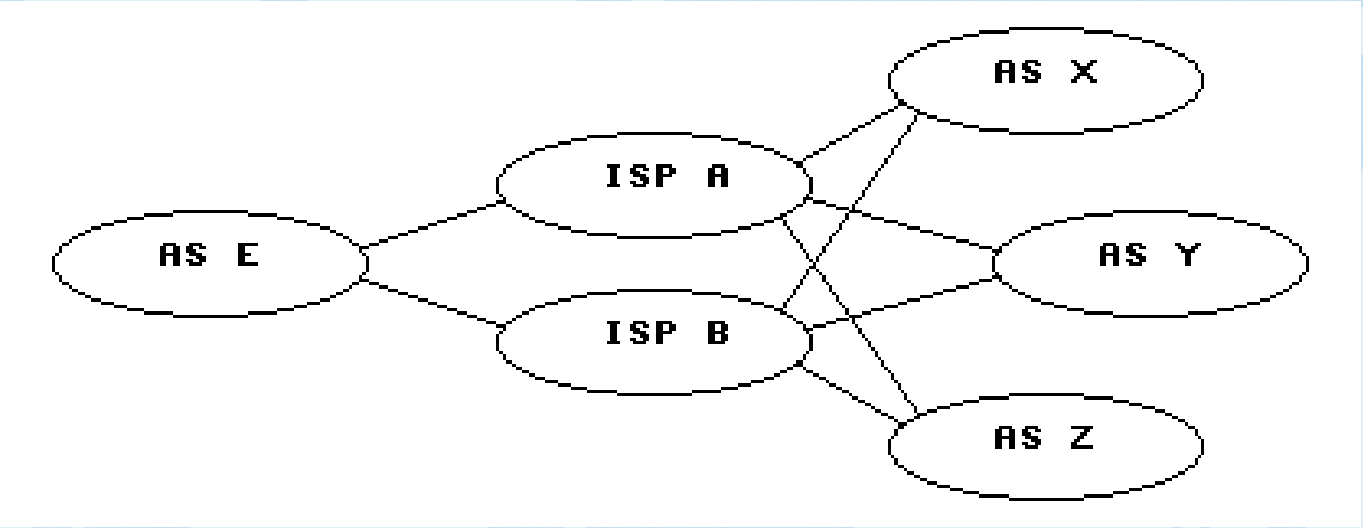
BGP wat kan je ermee

- Bepaalde routes uitfilteren:
  - Op prefix (bijvoorbeeld van klant: alleen vooraf opgegeven IP adresreeksen toegestaan)
  - Op AS pad (bijvoorbeeld naar peer: alleen routes met lokale AS en eventueel klant-ASen)
- Bepaalde routes altijd prefereren, bijvoorbeeld door routes van internet exchange een hogere local preference te geven, dit is immers goedkoper dan transit

Verkeer balanceren

- Bij meerdere verbindingen naar buiten (zonder dat heeft BGP weinig nut...) balanceert BGP automatisch het verkeer
- Helaas niet altijd zoals je wilt
- Zelf beinvloeden:
  - AS pad langer maken voor bepaalde routes, dit maakt ze minder aantrekkelijk
  - MED instellen om doorslag te geven wanneer AS pad even lang is

Soms te effectief



- In het geval van twee ISPs die met dezelfde netwerken peeren zullen veel paden even lang zijn
- Iedere aanpassing heeft gelijk enorme impact op de verkeersverdeling

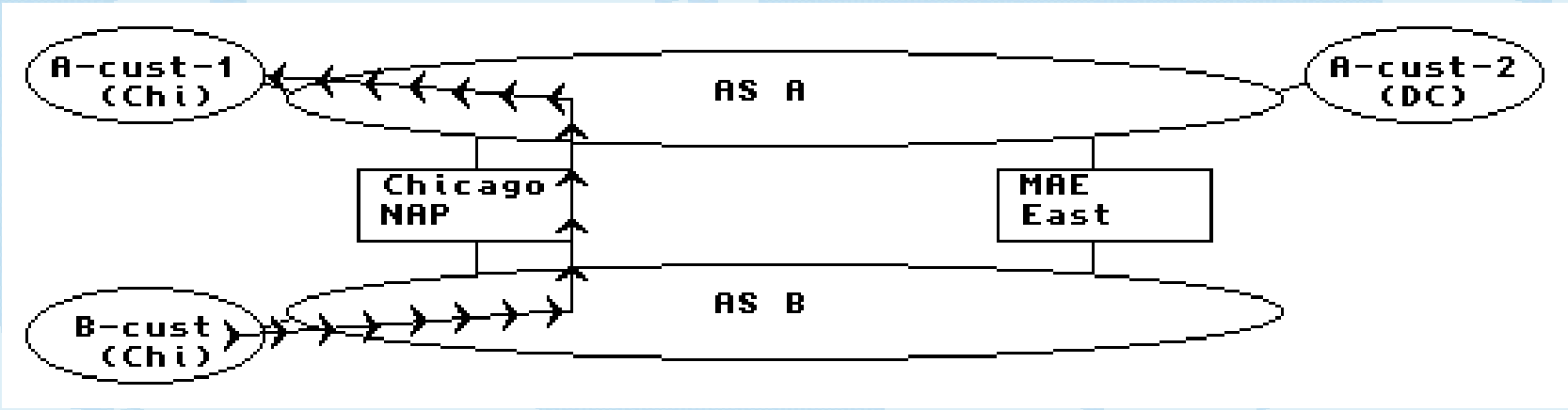
Communities praktisch

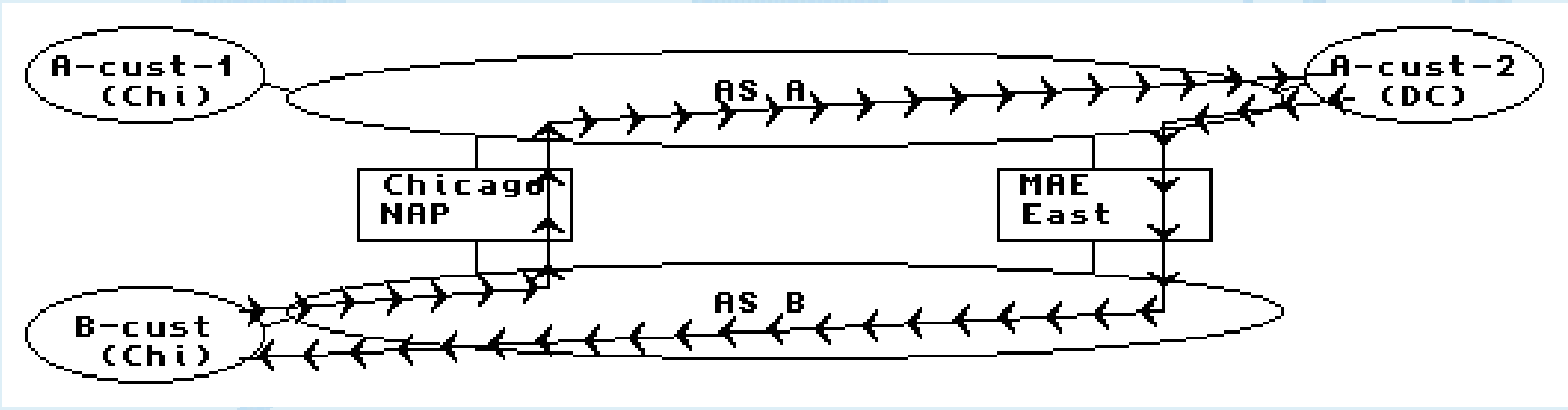
```
Aut-num:      AS702
as-name:      AS702
descr:        UUNET - Commercial IP service provider in Europe
import:       from AS72 194.98.169.195 at 194.98.169.196 accept AS72
import:       from AS109 213.53.49.50 at 213.53.49.49 accept AS109
[...]
export:       to AS72 194.98.169.195 at 194.98.169.196 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:      ------------------------------------------------------
```

Hoe gebruiken ISPs BGP

- Filteren, filteren, filteren
- Nouja, niet iedereen...
- Kleine adresblokken worden uitgefilterd:
  120.000 prefixen is meer dan genoeg!
- "Hot potato" of "early exit" routing

Early exit: simpel





Multiprotocol BGP

- Extensie op BGP4 die het mogelijk maakt routing informatie voor andere protocollen door te geven
- Gebruikt voor multicast: het versturen van een enkel pakket naar meerdere ontvangers
- En ook IPv6: de nieuwe versie van het IP protocol die veel meer adressen mogelijk maakt

Beveiliging van BGP

- Gevoelig op TCP, IP en ethernet-niveau
- Maar uitbuiten zeker niet makkelijk! En: MD5 / wachtwoord, maar niet al te sterk en lastig
- Toekomst:
  - S-BGP: Secure BGP
  - soBGP: secure origin BGP
- Zware encryptie teveel van het goede? Overgrote meerderheid problemen door fouten

Risicos huidige werkwijze

- Filtering maakt inspelen op veranderingen moeilijk
- Alles komt op maar enkele punten bij elkaar, zie New York op 11 september
- Extra verbindingen blijken in praktijk vaak slechter te werken dan de bedoeling
- 120.000 routes is VEEL, weinig foutmarge
- Faillisementen helpen niet bepaald

Dat was het dan.

Bedankt voor de aandacht!

Iljitsch van Beijnum

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

:-)