

Distributed Hash Table (DHT)

DHT = distributed P2P database

Database has (key, value) pairs;

- key: ss number; value: human name
- key: content type; value: IP address

Peers **query** DB with key

- DB returns values that match the key

Peers can also **insert** (key, value) pairs

1

DHT Identifiers

Assign integer identifier to each peer in range $[0, 2^n - 1]$.

- Each identifier can be represented by n bits.

Require each key to be an integer in **same range**.

To get integer keys, hash original key.

- eg, key = $h(\text{"Led Zeppelin IV"})$
- This is why they call it a distributed "hash" table

2

How to assign keys to peers?

Central issue:

- Assigning (key, value) pairs to peers.

Rule: assign key to the peer that has the **closest** ID.

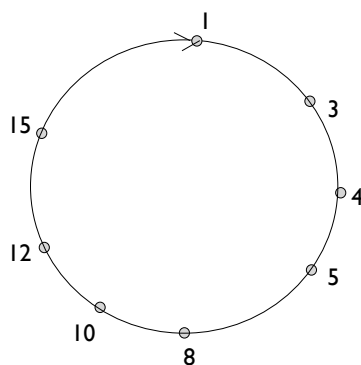
Convention in lecture: closest is the **immediate successor** of the key.

Ex: $n=4$; peers: 1,3,4,5,8,10,12,14;

- key = 13, then successor peer = 14
- key = 15, then successor peer = 1

3

Circular DHT (1)



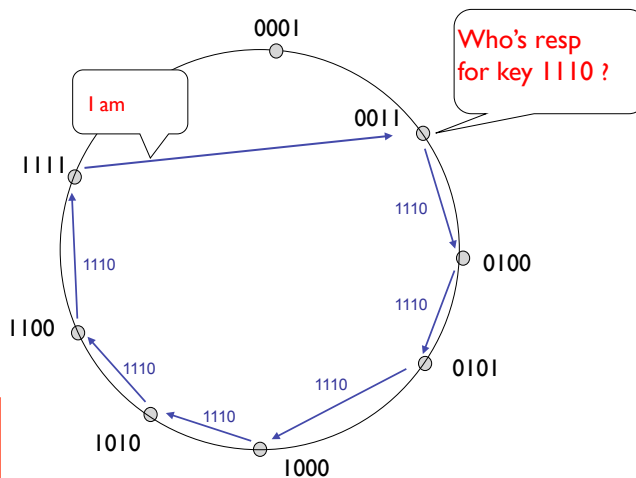
Each peer *only* aware of immediate successor and predecessor.
“Overlay network”

4

Circle DHT (2)

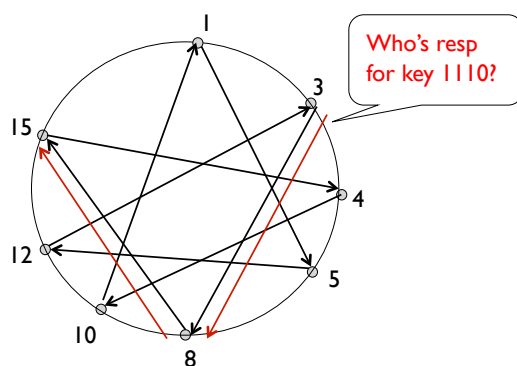
$O(N)$ messages
on avg to resolve
query, when there
are N peers

Define closest
as closest
successor



5

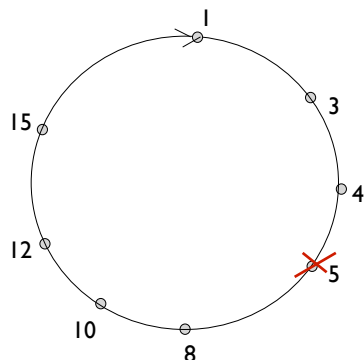
Circular DHT with Shortcuts



Each peer keeps track of IP addresses of predecessor, successor, short cuts.
Reduced from 6 to 2 messages.
Possible to design shortcuts so $O(\log N)$ neighbors, $O(\log N)$ messages in query

6

Peer Churn



- To handle peer churn, require each peer to know the IP address of its two successors.
- Each peer periodically pings its two successors to see if they are still alive.

Peer 5 abruptly leaves

Peer 4 detects; makes 8 its immediate successor; asks 8 who its immediate successor is; makes 8's immediate successor its second successor.

What if peer 13 wants to join?

7

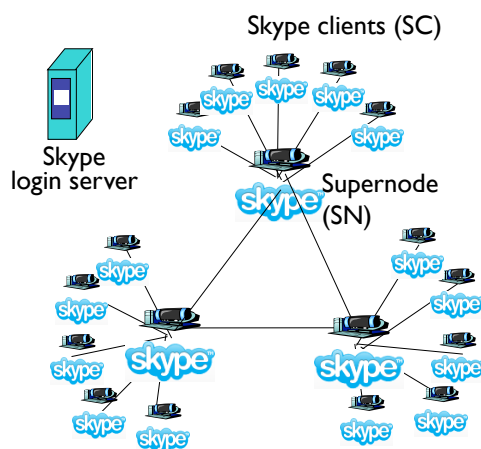
P2P Case study: Skype

inherently P2P: pairs of users communicate.

proprietary application-layer protocol (inferred via reverse engineering)

hierarchical overlay with SNs

Index maps usernames to IP addresses; distributed over SNs



8

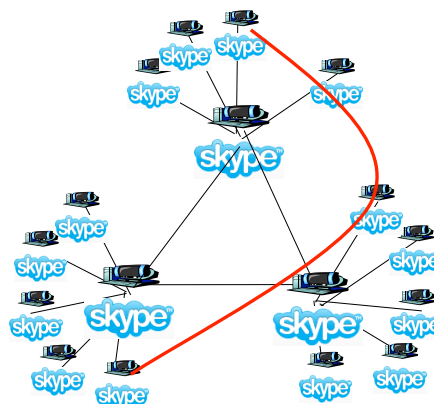
Peers as relays

Problem when both Alice and Bob are behind “NATs”.

- NAT prevents an outside peer from initiating a call to insider peer

Solution:

- Using Alice's and Bob's SNs, Relay is chosen
- Each peer initiates session with relay.
- Peers can now communicate through NATs via relay



9

Routing on the Internet

In the beginning there was the ARPANET:

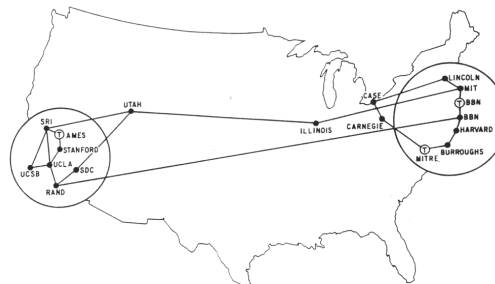
- route using GGP (Gateway-to-Gateway Protocol), a distance vector routing protocol

Problems:

- needed “flag-hour” to update routing protocol
- incompatibility across vendors

Solution: hierarchical routing

- administrative autonomy:
- each network admin can control routing within its own network
- internet: network of networks
- allows the Internet to scale:
 - with 200 million hosts, each router can't store all destinations in its routing table
 - route updates alone will swamp the links



MAP 4 September 1971

10

Hierarchical Routing

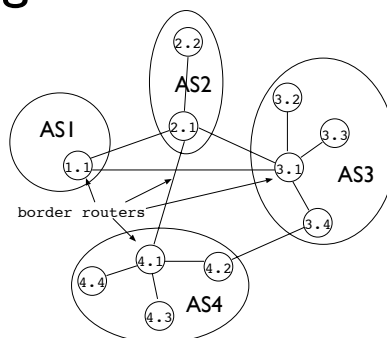
Aggregate routers into regions of “autonomous systems” (AS)

Routers in same AS run same routing protocol

- “intra-AS” routing protocol
- each AS uses its own link metric
- routers in different ASs can run different intra-AS routing protocol

Gateway/border router

- direct link to router in other AS(s)
- keeps in its routing table:
 - next hop to other ASs
 - all hosts within its AS
- hosts within an AS only keep a default route to the border router



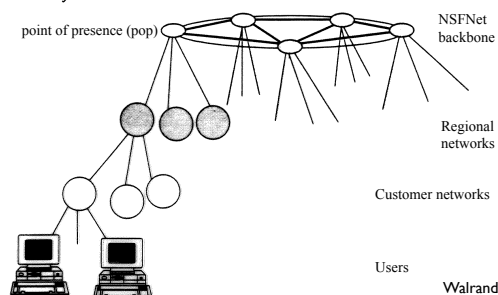
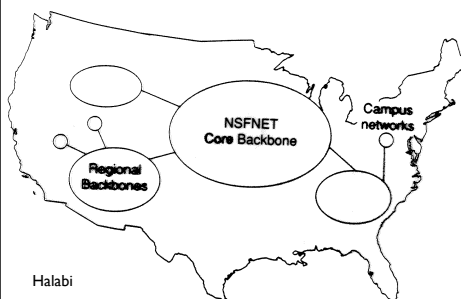
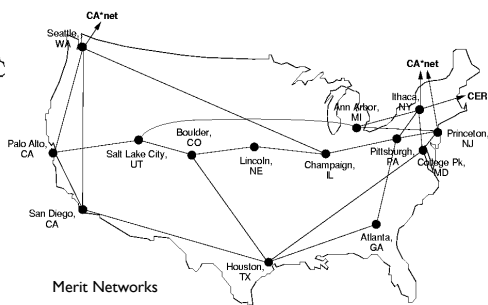
3.1	
dest	next
1.*	1.1
2.*	2.1
4.*	2.1
3.2	3.2
3.3	3.3
3.4	3.4

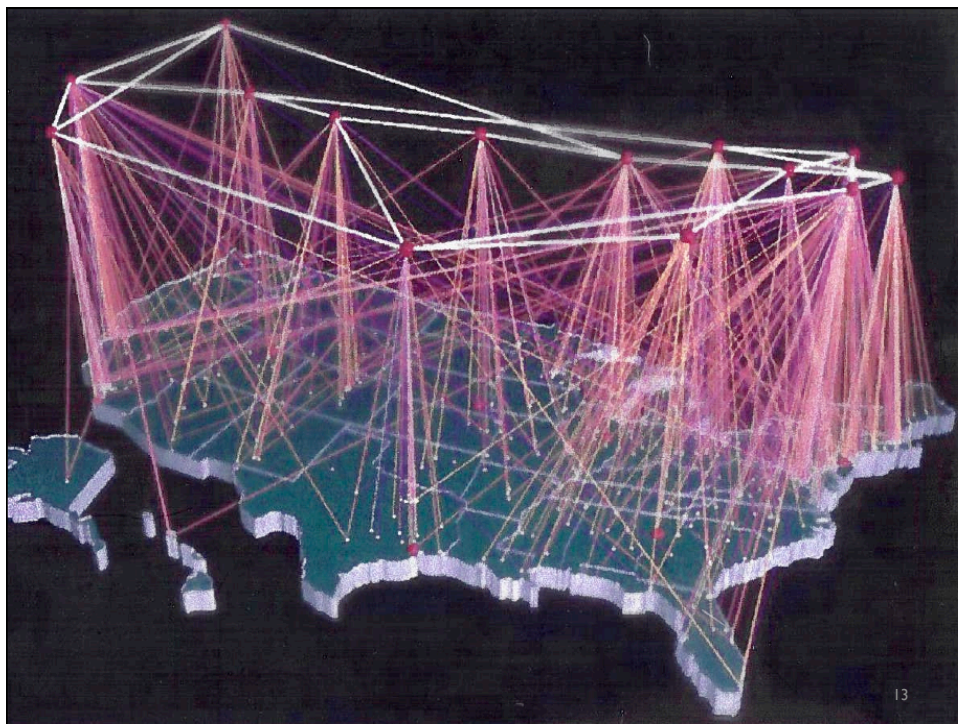
The NSFNet 1989

Area hierarchy:

- backbone/core: NSFNet
- regional networks: MichNet, BARRNET, Los Nettos, Cerfnet, JVCNet, NEARNet, etc.
- campus networks

Sketch of the Internet around 1990.



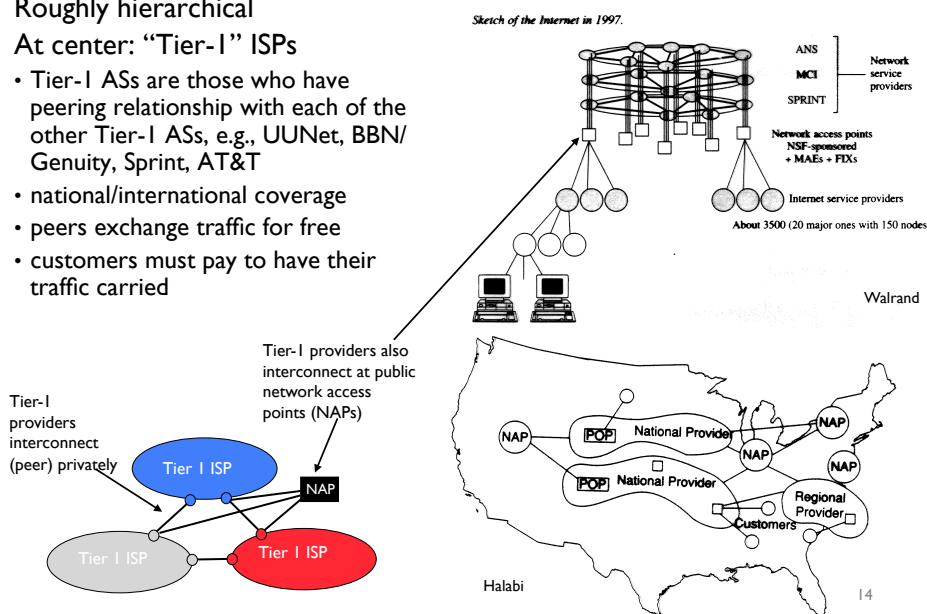


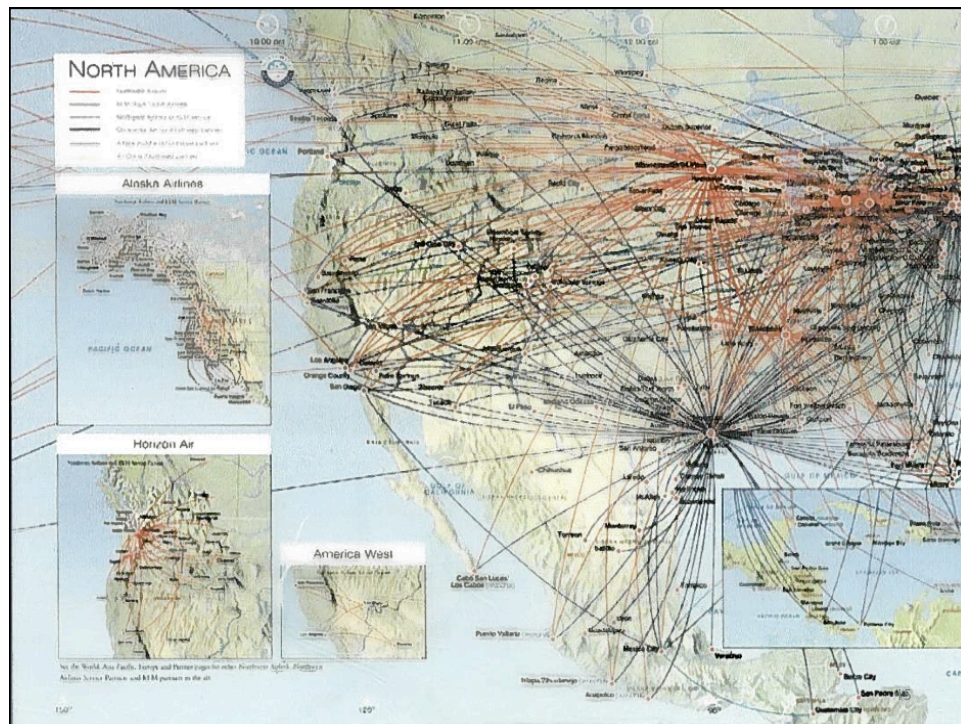
Commercialization (1994)

Roughly hierarchical

At center: "Tier-I" ISPs

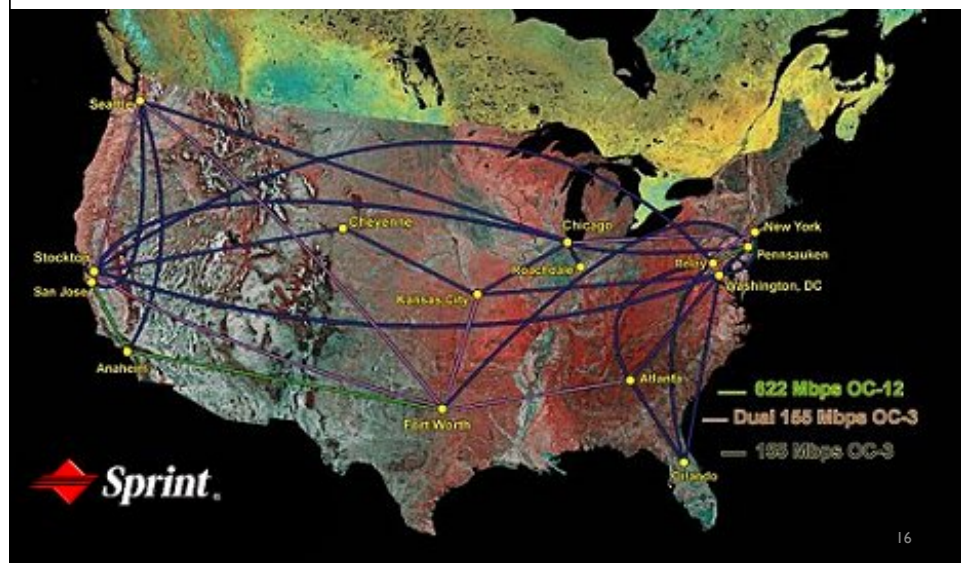
- Tier-I ASs are those who have peering relationship with each of the other Tier-I ASs, e.g., UUNet, BBN/Genuity, Sprint, AT&T
- national/international coverage
- peers exchange traffic for free
- customers must pay to have their traffic carried



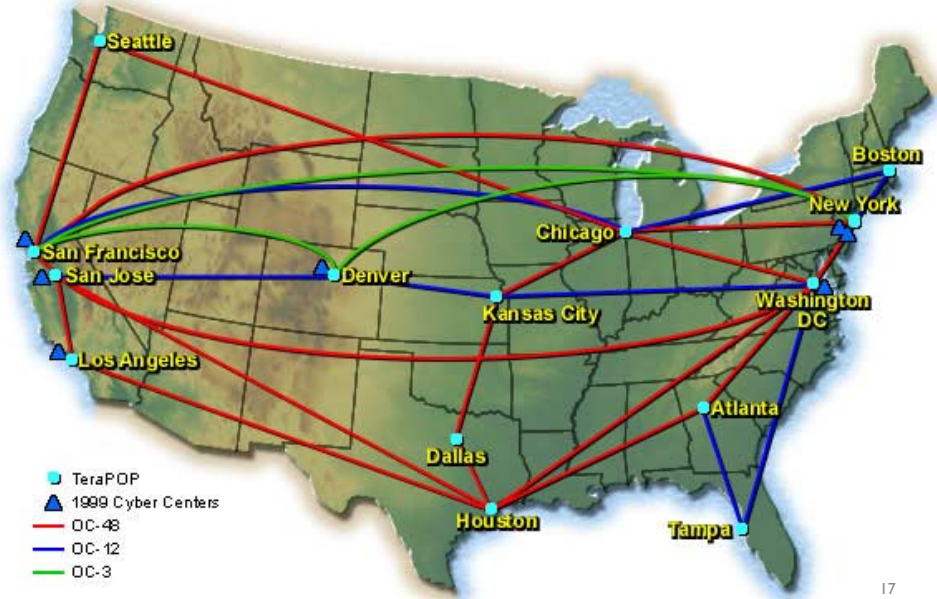


Tier-I ISP: e.g., Sprint

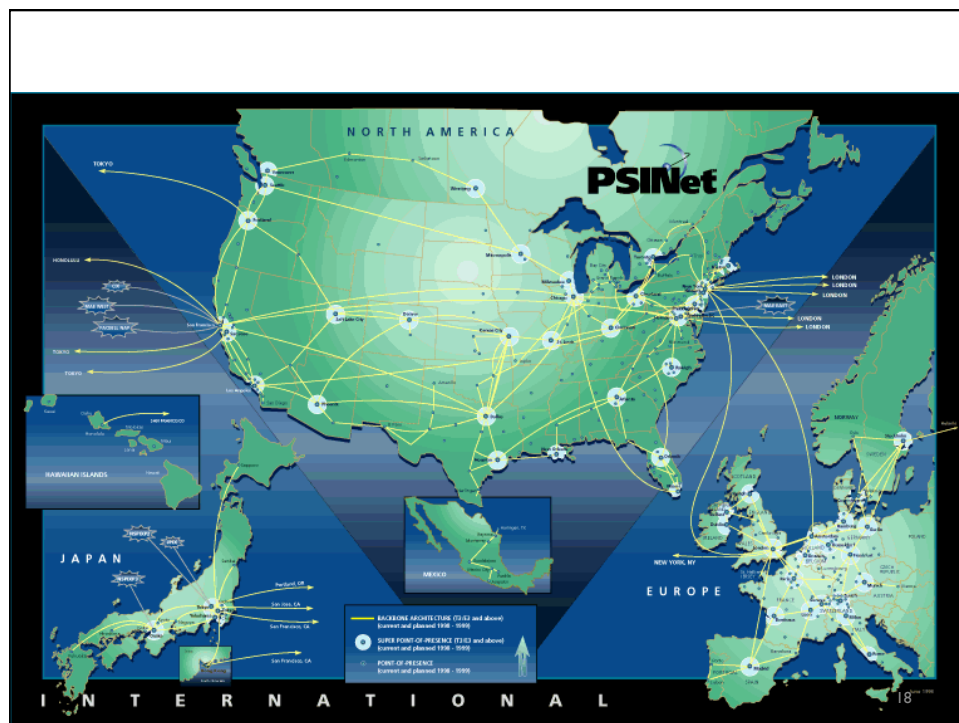
Sprint US backbone network

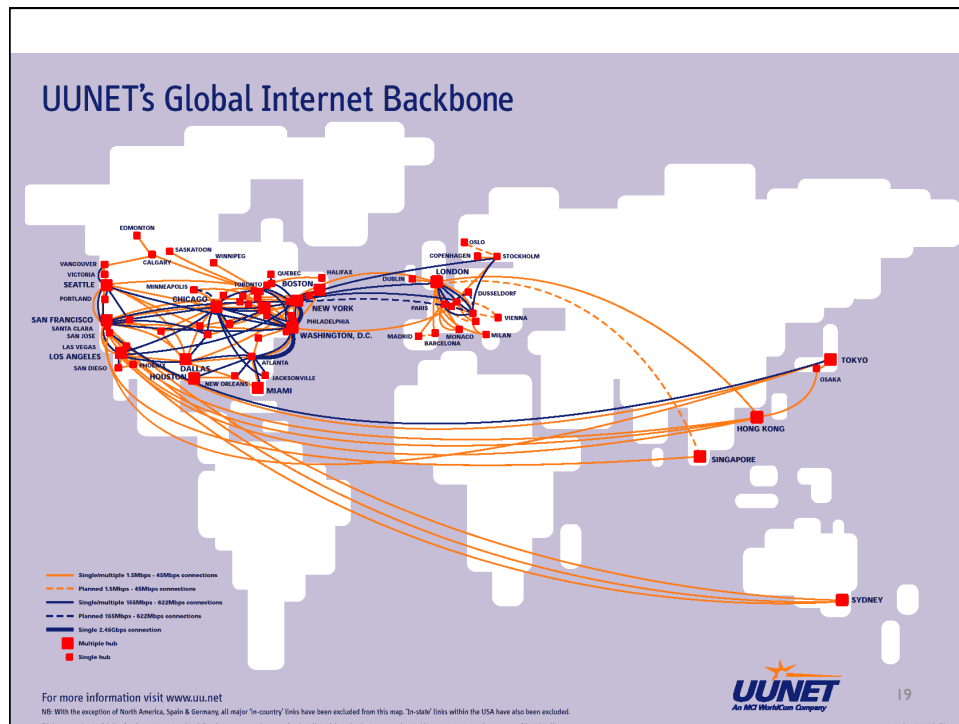


Qwest IP Backbone (Late 1999)



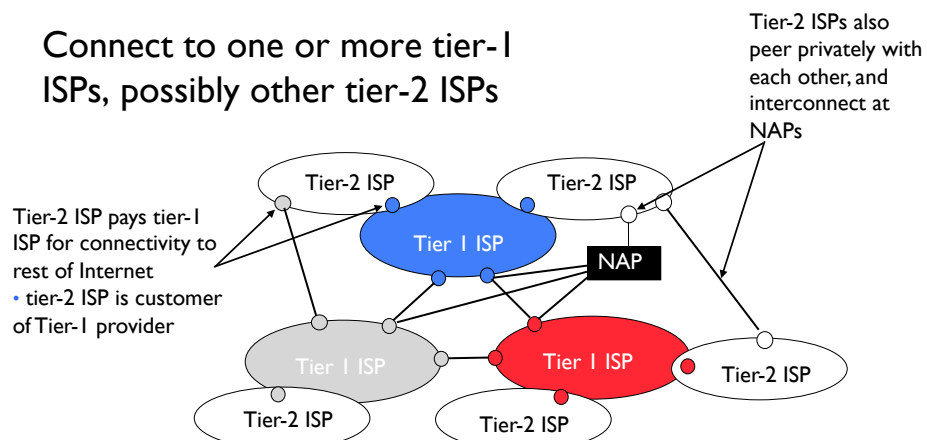
17





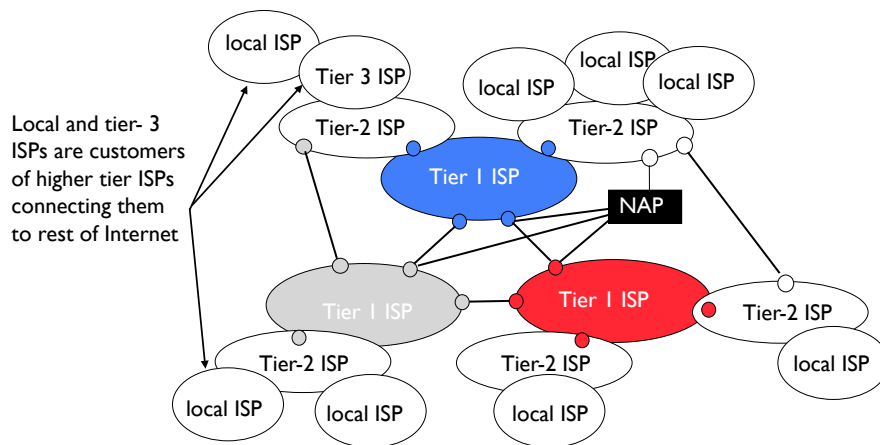
“Tier-2” ISPs: Smaller (Often Regional) ISPs

Connect to one or more tier-1
ISPs, possibly other tier-2 ISPs



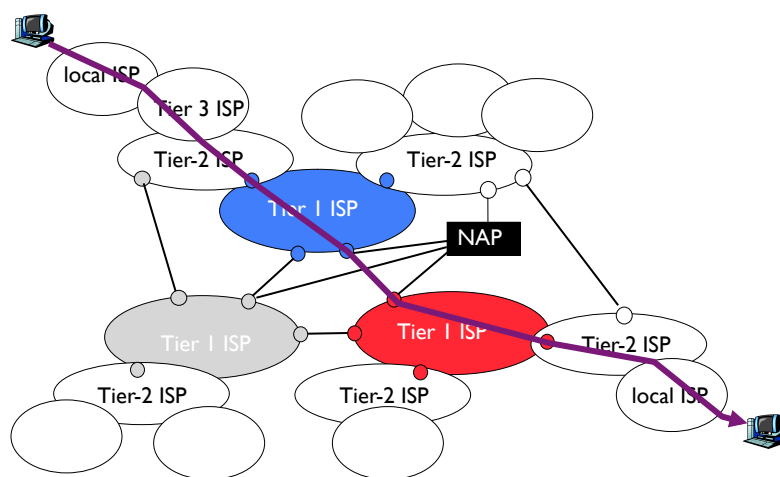
“Tier-3” ISPs and local ISPs

Last hop (“access”) network (closest to end systems)



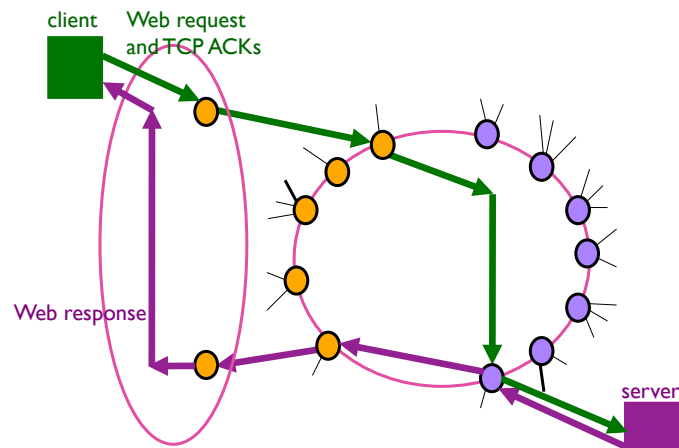
21

A Packet Passes Through Many Networks



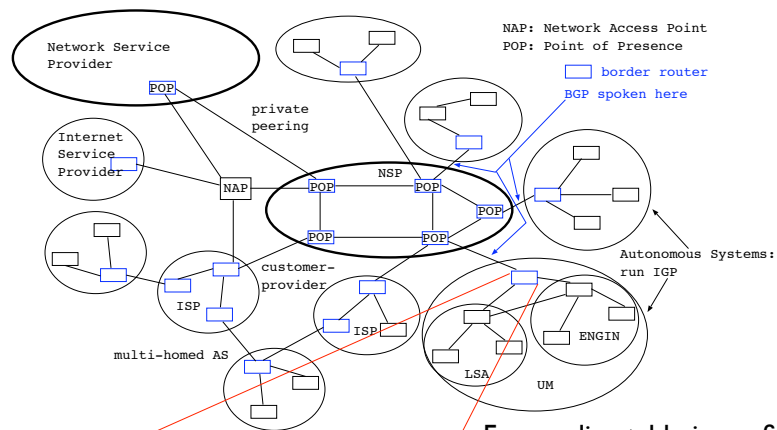
22

Routing is not Symmetric



23

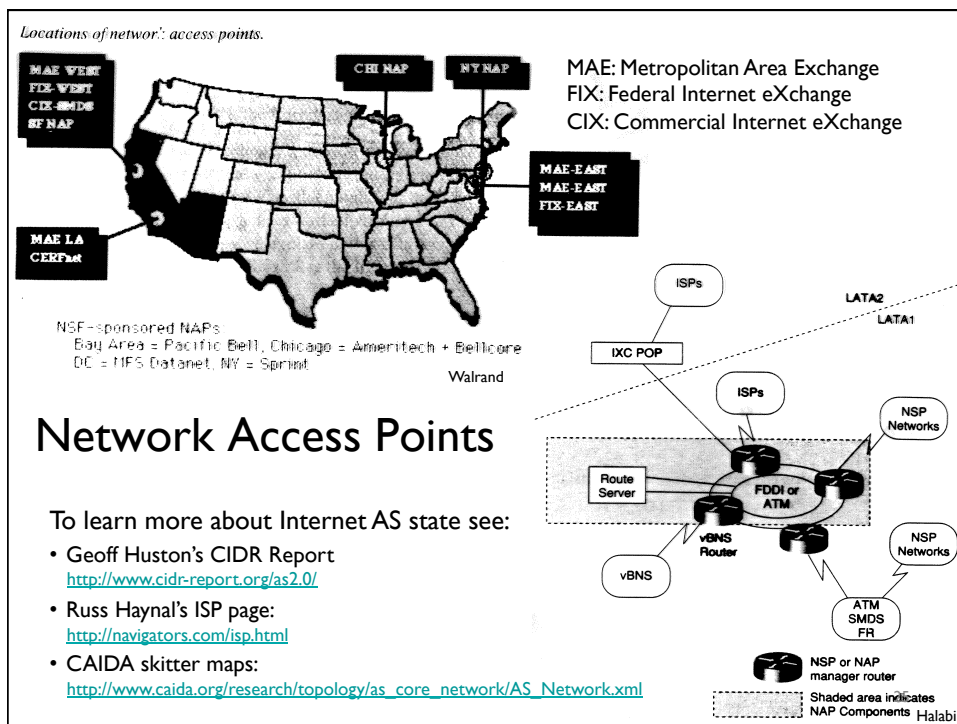
Interconnected ASes



Forwarding table is configured by both intra- and inter-AS routing algorithm

- Intra-AS sets entries for internal dests
- Inter-AS & Intra-As sets entries for external dests

24



ISP Backbone Links

Categories of links, by inside/outside AS:

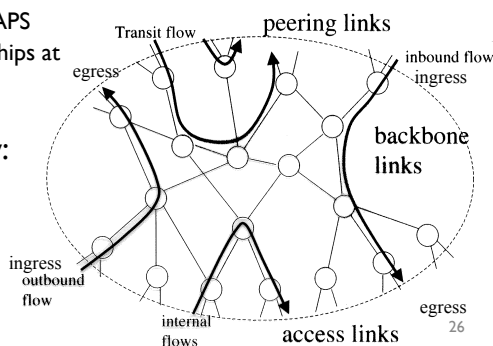
- backbone links: connect routers inside the backbone
- edge links: connect an AS with other ASs

Categorization by commercial relationship:

- access links: customers to ISP (can be multi-homed)
- peering links: inter-AS connections
 - either direct peering or through NAPS
 - may have multiple peering relationships at different geographic locations

Categorization by traffic flow:

- ingress link
- egress link



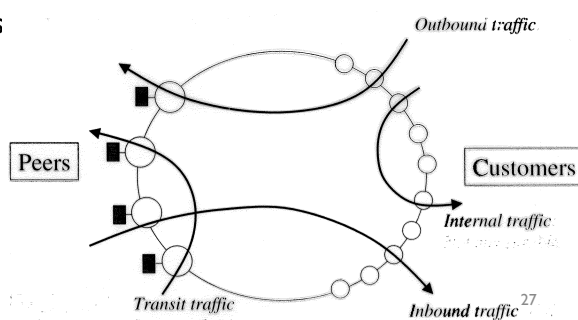
ISP Backbone Traffic

Types of traffic:

- internal traffic: between 2 access links
- transit traffic: between 2 peering links
- inbound traffic: ingress at peering link, egress at access link
- outbound traffic: vice versa

An ISP may choose to carry transit traffic

Most traffic on large ISPs



Internet inter-AS Routing: BGP

BGP (Border Gateway Protocol) is *the* de facto standard for inter-AS routing

- 06/89 v.1
- 06/90 v.2 EGP (Exterior Gateway Protocol) to BGP transition
- 10/91 v.3 BGP installed
- 07/94 v.4 *de facto* standard

BGP provides each AS a means to:

- obtain subnet (prefix) reachability information from neighboring ASs
- propagate the reachability information to all routers internal to the AS
- determine “good” routes to subnets based on reachability information **and policy**
- Inter-AS routing is **policy driven**, not load-sensitive, generally not QoS-based

Internet inter-AS Routing: BGP

When AS2 advertises a prefix to AS1, AS2 is *promising* AS1 to forward any datagrams destined to that prefix coming from AS1

- AS2 can aggregate prefixes in its advertisement

Pairs of routers (BGP peers) exchange reachability information over semi-permanent TCP connections: BGP sessions

- advantage of using TCP: reliable transmission allows for incremental updates
- disadvantage: TCP congestion control mechanism slows down route updates that could decongest link!
- BGP sessions do not correspond to physical links

29

Path Attributes & BGP Routes

When advertising a prefix, advertisement includes BGP attributes

Two important attributes:

- AS-PATH: the path vector of ASs through which the advertisement for a prefix passed through
- NEXT-HOP: the specific internal-AS router to next-hop AS (there may be multiple exits from current AS to next-hop-AS)

Sample BGP entry:

destination	NEXT-HOP	AS-PATH
198.32.163.0/24	202.232.1.8	2497 2914 3582 4600

- address range 198.32.163.0/24 is in AS 4600
- to get there, send to next hop router, which has address 202.232.1.8
- the path there goes through ASs 2497, 2914, 3582, in order

30

BGP Messages

BGP messages:

- **OPEN:** opens TCP connection to peer and authenticates sender
- **UPDATE:** advertises a new path (or withdraws an old one)
- **KEEPALIVE:** keeps connection alive in the absence of UPDATES; also acknowledges OPEN request
- **NOTIFICATION:** reports errors in previous message; also used to close connection

Internal Peering (iBGP):

- normally, external routes are not propagated within an AS
- to allow for propagation of path vectors, needs a BGP connection between two routers within an AS

31

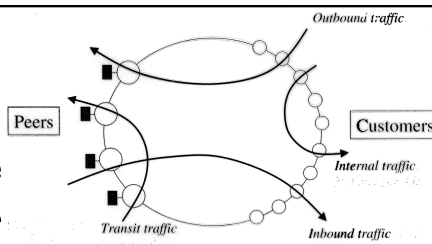
BGP Policy Routing

Some ASs are more equal than others
Commercial relationship between ASs

- peering: peers agree to exchange traffic for free
- customer-provider: customer pays provider for access
- backup

An AS's export policy (which routes it will advertise):

- to a customer: all routes
- to a peer or service provider: routes to all its own prefix address ranges and to its customers' prefixes, **not** to prefixes learned from other providers or peers
- internal routing of an AS is effected by its neighbors' route export policy



32

BGP Policy Tools

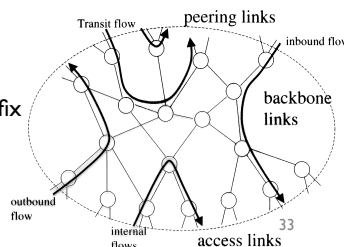
An AS may learn about more than one route to some prefix
Each AS applies its own local policies to select route

Import policies: which of the advertised routes I want to use

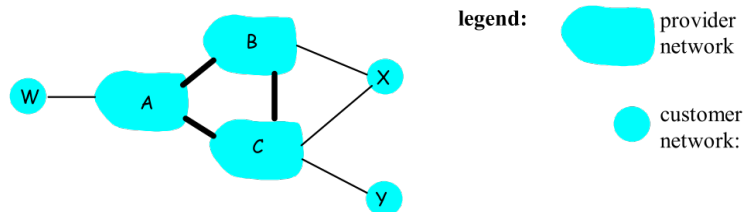
- always check AS-PATH against routing loop
- an AS can specify its preferred **egress** point to reach a specific other AS

Export policies: how to set attributes of routes I advertise

- always prepend itself to the AS-PATH
- multiple-exit discriminator (MED):
 - an AS can tell a neighbor its preferred **ingress** point
 - cold-potato routing: ingress closest to destination prefix
 - hot-potato routing: egress (NEXT-HOP) closest to traffic source (ignore the other guy's MED)

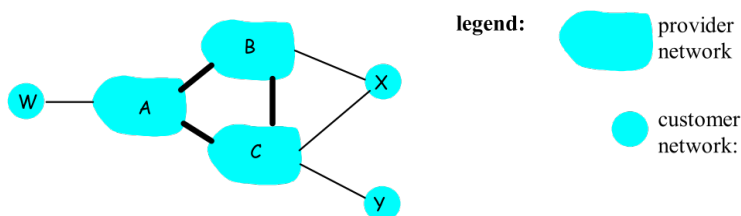


BGP Routing Policy Example



A,B,C are provider networks
X,W,Y are customer (of provider networks)
X is dual-homed: attached to two networks
X does not want to route from B via X to C
.. so X will not advertise to B a route to C

BGP Routing Policy Example



A advertises to B the path AW
 B advertises to X the path BAW
 Should B advertise to C the path BAW?
 No way! B gets no “revenue” for routing
 CBAW since neither W nor C are B’s customers
 B wants to force C to route to w via A
 B wants to route *only* to/from its customers!

35

Why Separate Intra- and Inter-AS Routing ?

Policy:

Inter-AS: admin wants control over how its traffic is routed, who routes through its network. Inter-AS routing is policy driven, not load-sensitive, generally not QoS-based

Intra-AS: single admin, so no policy decisions needed

Scale:

hierarchical routing saves table size, reduced update traffic

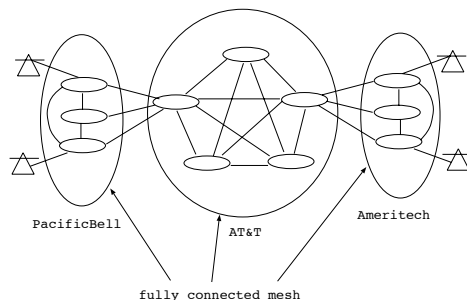
Performance:

Intra-AS: can focus on performance

Inter-AS: policy may dominate over performance

36

The Phone Network (for Contrast)



Phone routing:

- within central office, connect directly (local call)
- between central offices (zones), use 1 hop path
- else, send to core (long distance)
- core: use 1-hop path first, else try alternate 2-hop paths

(old) AT&T:

135 switches, 2.5 million trunks
fully connected networks

Phone network downtime:

- switches: a few mins/year
- links: a few hours/year

The Internet: frequent outages, each lasting from a few minutes to a few days

37

Internet Protocol Stack

application: supporting network applications

- HTTP, SMTP, FTP, etc.

transport: endhost-endhost data transfer

- TCP, UDP

network: routing of datagrams from source to destination

- IP, routing protocols

link: data transfer between neighboring network elements

- Ethernet, WiFi

physical: bits “on the wire”



38

Physical Layer: Signals

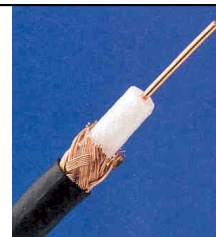
We only look at a very brief overview of the physical layer in this course

- to learn more, take EECS 455: Signals and Systems, EECS 554: Digital Communication and Coding, and/or EECS 557: Communication Networks

Signal degrades (attenuates) as it travels further from the source (caused by resistance on the wire, cosmic interference, etc.)

Wires:

- coax: shielding of core reduces interference
- twisted pair: twisting a pair of wires changes the electrical property of the pair, reducing interference
- glass fiber:
 - LED or laser as signal source
 - more fragile but no interference
 - can carry more data
 - hard to splice



Shielded twisted pair (STP)



Unshielded twisted pair (UTP)



39

Ethernet Wiring

Network interface cards (NICs)

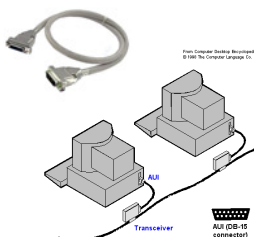
Ethernet card



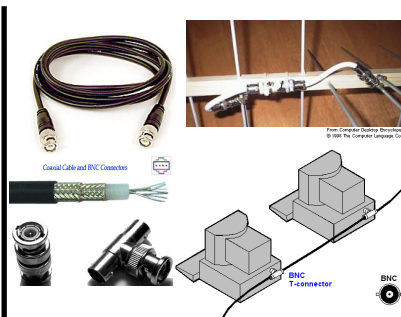
Wireless card



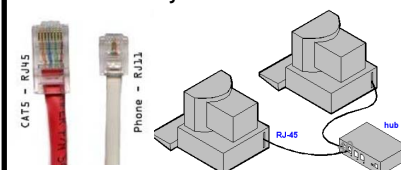
Thick (10Base5) Ethernet with AUI connector



Thin (10Base2) Ethernet with BNC connector



Twisted Pair (10/100BaseT/Gigabit) Ethernet with RJ-45 connector

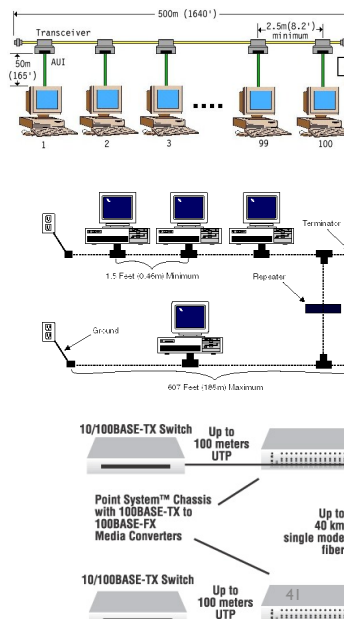


Transmission Distance Limitation

Transmission distance limited due to: *signal loss* caused by *interference* and *sharing condition*

Examples:

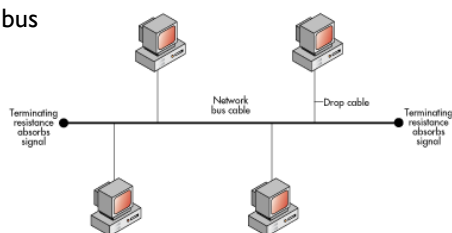
- serial line (RS-232): 15 m
- 10Base5 Ethernet: 500 m, at most 4 repeaters (2.5km total)
- 10Base2 Ethernet: 185 m max, 0.45m min, at most 4 repeaters (1 km total)
- 10/100BaseT Ethernet (copper): 100 m
- 100 Mbps Ethernet (fiber): 412 m
- fiber: 2-100 km



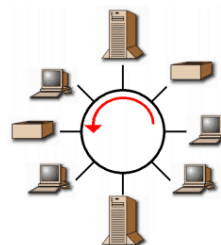
Wiring Scheme

Point-to-point: $\binom{N}{2} = \frac{N^2 - N}{2}$ connections to connect N computers

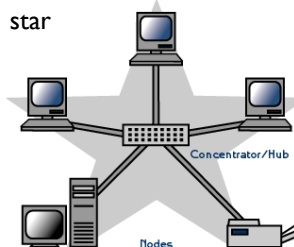
Shared LAN:
bus



ring



star



42

Wireless

Radio:

- satellite: order of Gbps, up-down latency of 250 ms (too long)
- cellular, WiFi, WiMax
- Bluetooth: 2.4 GHz short range radio, 721 Kbps – 2.1 Mbps, 1-100 m
- UWB: 3.1 GHz-10.6 GHz, 480-675 Mbps, 10 m, less interference due to use of short pulses
 - Wireless USB
 - Bluetooth 3.0
 - Wireless FireWire

Microwave:

- high bandwidth: 1.5 Gbps
- can be aimed in a single direction
- requires “line-of-sight”
- most useful to connect buildings on campus

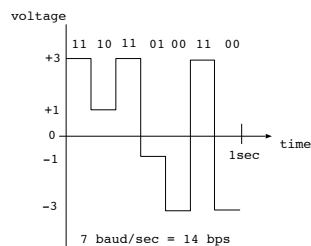
Infrared: shorter distance, no need for antenna

43

Signals

Bit vs. Baud

- bit transmitted as electrical or optical signal
- bit rate: number of **bits** per second
- baud rate: signal/voltage **level changes** per second
- each level can represent multiple bits
 - for binary signalling, bit rate == baud rate
 - for *M*-ary signalling, bit rate $\neq$ baud rate
 - example: 4-ary signalling

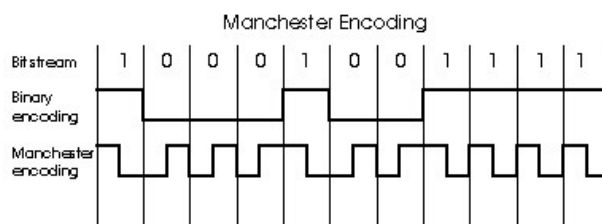


Example: RS-232

- negative voltage (-15V) represents a 1
- positive voltage (+15V) represents a 0
- bit rate == baud rate
- 7 bits/character
- to allow asynchronous communication:
 - 1 start bit, 1 stop bit

44

Manchester Encoding



Used in 10BaseT Ethernet

Each bit has a transition

Allows clocks in sending and receiving nodes to synchronize to each other

- no need for a centralized, global clock among nodes!

45

Signal Digitization

The maximum rate at which you can transmit data is limited by how fast (in Hertz) the sender's hardware can change voltage level and how sensitive the receiver's hardware is to voltage level changes

Nyquist Sampling Theorem (1924):

For a signal band limited in frequency at B Hz, we need to sample it at $2B$ Hz to be able to reconstruct the original signal from the samples

Conversely, and more generally, for an M -level signal, the maximum data rate (R) is determined by:

$$R = 2B \log_2 M \text{ bps,}$$

where B is the line bandwidth (in Hz)

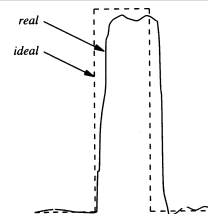
Example: RS-232, $M = 2$, phone line: $B = 3$ kHz

So, signal travelling over phone line using RS-232 signaling has a maximum data rate of $R = 2B = 6$ Kbps (modems don't use RS-232 signaling!)

46

Signal to Noise Ratio (dB)

Nyquist Sampling Theorem assumes noiseless channel
In reality, channels are noisy



S/N : ratio of signal power (watts) to noise power (watts) usually given as signal-to-noise ratio in quantity of $10 \log_{10} S/N$, called dB (decibels)

Examples:

- $S/N = 10$, signal-to-noise ratio is 10 dB
- $S/N = 100$, signal-to-noise ratio is 20 dB

Shannon Capacity (1948): the maximum data rate (C) of a noisy channel with bandwidth B Hz and a given signal-to-noise ratio is:

$$C = B \log_2 (1 + S/N) \text{ bps,}$$

which gives M on the order of $\sqrt{1 + S/N}$

Phone lines have $B = 3$ kHz, $S/N = 30$ db, so $C = 29.9$ kbps ($3 \cdot \log_2(1+1000)$)

How can one have 56 Kbps modem?

47

Carrier Wave

Observation: a continuous, oscillating signal propagates further (with less signal loss) than other signals

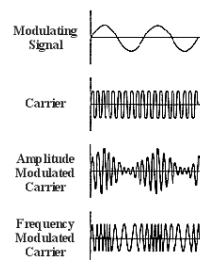
Hence to send data long distance, we use a continuous sine wave as a *carrier wave*

Data is “carried” by modifying the carrier wave, a process called **modulation**

Two types of modulation:

1. Amplitude Modulation (AM): not as robust
2. Frequency Modulation (FM): more robust

Modem: modulator-demodulator



48

Transmission Bandwidth (2009)

Backbone link technologies

OC-768, STM-256x	40 Gbps
OC-192, STM-64x	10 Gbps
OC-48, STS-48	2.5 Gbps (Tier-1 ISPs)
OC-12, STS-12	622 Mbps (Tier-2 ISPs)
OC-3, STS-3, STM-1	155 Mbps
OC-1 (Sonet), STS-1	51 Mbps
T-3, DS-3 N. America	44 Mbps
E-3 Europe	34 Mbps
E-2 Europe	8 Mbps
T-2, DS-2 N. America	6 Mbps
E-1, DS-1 Europe	2 Mbps
T-1, DS-1 N. America	1.5 Mbps
DS-0, PCM	64 Kbps

Access link technologies

cable modem (down)	30 Mbps
ADSL2+	24 Mbps/1 Mbps
ADSL2	12 Mbps/1 Mbps
cable modem (down)	10 Mbps
Standard ADSL	8 Mbps/1 Mbps
ADSL Lite	1.5 Mbps/512 Kbps
ISDN	128 Kbps
U.S. Robotics x2	56 Kbps
56flex56flex, x2	33.6 Kbps
V.34, Rockwell V.Fast	28.8 Kbps
V.32bis modem, V.17 fax	14.4 Kbps
speed circa early 1990s	9600 bps
speed circa 1980s	1200-2400 bps

49

Transmission Bandwidth (2009)

LAN technologies wired

10GigEthernet	10 Gbps
Gigabit Ethernet	1 Gbps
FDDI, Fast Ethernet	100 Mbps
Token Ring	16 Mbps
Ethernet	10 Mbps

wireless

WiFi 802.11a	23 Mbps
WiFi 802.11g	19 Mbps
WiFi 802.11b	4.5 Mbps

PAN technologies

USB3	5 Gbps
UWB: Bluetooth3, Wireless USB, Wireless FireWire	480 Mbps
USB2	480 Mbps
USB1	12 Mbps
Bluetooth2	2 Mbps
Bluetooth1	723 Kbps

Cable bandwidth

Cat-5 cable	100 Mbps
Cat-4 cable	20 Mbps
Cat-3 cable	10 Mbps
Level 1 cable (min)	20 Kbps

Wide-area wireless data technologies

WiMax 802.16	75 Mbps
WiBro (SKorea, mobile WiMax) 802.16e	30-50 Mbps
CDMA 1xEVDO rev. B	5 Mbps
HSDPA	3.6 Mbps
CDMA 1xEVDO rev.A	3 Mbps
EDGE	236 Kbps
CDMA 1x	163 Kbps
GPRS	80 Kbps

50