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

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
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

Circular DHT (1)

Each peer only aware of immediate successor and predecessor.
“Overlay network”
Circle DHT (2)

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

Define closest as closest successor.

Who’s resp for key 1110?

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.

Who’s resp for key 1110?
Peer Churn

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?

• 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.

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
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

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
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

<table>
<thead>
<tr>
<th></th>
<th>dest</th>
<th>next</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.4</td>
<td></td>
</tr>
</tbody>
</table>

The NSFNet 1989

Area hierarchy:
- backbone/core: NSFNet
- regional networks: MichNet, BARRNET, Los Nettos, Cerfnet, JVCNet, NEARNet, etc.
- campus networks
Commercialization (1994)

Roughly hierarchical

At center: “Tier-1” ISPs

- Tier-1 ASs are those who have peering relationship with each of the other Tier-1 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-1 ISP: e.g., Sprint

Sprint US backbone network
“Tier-2” ISPs: Smaller (Often Regional) ISPs

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

Tier-2 ISP pays tier-1 ISP for connectivity to rest of Internet
- tier-2 ISP is customer of Tier-1 provider

Tier-2 ISPs also peer privately with each other, and interconnect at NAPs
“Tier-3” ISPs and local ISPs

Last hop ("access") network (closest to end systems)

Local and tier-3 ISPs are customers of higher tier ISPs connecting them to rest of Internet

A Packet Passes Through Many Networks
Routing is not Symmetric

Intra-AS Routing algorithm
- Sets entries for internal dests

Inter-AS Routing algorithm
- Sets entries for external dests

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
Network Access Points

To learn more about Internet AS state see:

- Geoff Huston’s CIDR Report
  http://www.cidr-report.org/as2.0/
- Russ Haynal’s ISP page:
  http://navigators.com/isp.html
- CAIDA skitter maps:
  http://www.caida.org/research/topology/as_core_network/AS_Network.xml

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

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:

<table>
<thead>
<tr>
<th>destination</th>
<th>NEXT-HOP</th>
<th>AS-PATH</th>
</tr>
</thead>
<tbody>
<tr>
<td>198.32.163.0/24</td>
<td>202.232.1.8</td>
<td>2497 2914 3582 4600</td>
</tr>
</tbody>
</table>

- 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
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

BGP Policy Routing

Some ASs are more equal than other
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
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!

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
The Phone Network (for Contrast)

(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

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

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

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”
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

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 \( \frac{N}{2} \) connections to connect \( N \) computers

Shared LAN:
- bus
- ring
- star

\[ \text{Connections} = \frac{N^2 - N}{2} \]
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
- UVWB: 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

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 ≠ 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
**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!

---

**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!)
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 (decibles)
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?

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
### Transmission Bandwidth (2009)

#### Backbone link technologies

<table>
<thead>
<tr>
<th>Technology</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC-768, STM-256x</td>
<td>40 Gbps</td>
</tr>
<tr>
<td>OC-192, STM-64x</td>
<td>10 Gbps</td>
</tr>
<tr>
<td>OC-48, STS-48</td>
<td>2.5 Gbps</td>
</tr>
<tr>
<td>OC-12, STS-12</td>
<td>622 Mbps</td>
</tr>
<tr>
<td>OC-3, STS-3, STM-1</td>
<td>155 Mbps</td>
</tr>
<tr>
<td>OC-1 (Sonet), STS-1</td>
<td>51 Mbps</td>
</tr>
<tr>
<td>T3, DS-3 N. America</td>
<td>44 Mbps</td>
</tr>
<tr>
<td>E-3 Europe</td>
<td>34 Mbps</td>
</tr>
<tr>
<td>E-2 Europe</td>
<td>8 Mbps</td>
</tr>
<tr>
<td>T2, DS-2 N. America</td>
<td>6 Mbps</td>
</tr>
<tr>
<td>E-1, DS-1 Europe</td>
<td>2 Mbps</td>
</tr>
<tr>
<td>T-1, DS-1 N. America</td>
<td>1.5 Mbps</td>
</tr>
<tr>
<td>DS-0, PCM</td>
<td>64 Kbps</td>
</tr>
</tbody>
</table>

#### Access link technologies

<table>
<thead>
<tr>
<th>Technology</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>cable modem (down)</td>
<td>30 Mbps</td>
</tr>
<tr>
<td>ADSL2+</td>
<td>24 Mbps/1 Mbps</td>
</tr>
<tr>
<td>ADSL2</td>
<td>12 Mbps/1 Mbps</td>
</tr>
<tr>
<td>cable modem (down)</td>
<td>10 Mbps</td>
</tr>
<tr>
<td>Standard ADSL</td>
<td>8 Mbps/1 Mbps</td>
</tr>
<tr>
<td>ADSL Lite</td>
<td>1.5 Mbps/512 Kbps</td>
</tr>
<tr>
<td>ISDN</td>
<td>128 Kbps</td>
</tr>
<tr>
<td>U.S. Robotics x2</td>
<td>56 Kbps</td>
</tr>
<tr>
<td>56flex56flex, x2</td>
<td>33.6 Kbps</td>
</tr>
<tr>
<td>V.34, Rockwell VFast</td>
<td>28.8 Kbps</td>
</tr>
<tr>
<td>V.32bis modem, V.17 fax</td>
<td>14.4 Kbps</td>
</tr>
<tr>
<td>speed circa 1990s</td>
<td>9600 bps</td>
</tr>
<tr>
<td>speed circa 1980s</td>
<td>1200-2400 bps</td>
</tr>
</tbody>
</table>

### LAN technologies

#### Wired

<table>
<thead>
<tr>
<th>Technology</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>10GigEthernet</td>
<td>10 Gbps</td>
</tr>
<tr>
<td>Gigabit Ethernet</td>
<td>1 Gbps</td>
</tr>
<tr>
<td>FDDI, Fast Ethernet</td>
<td>100 Mbps</td>
</tr>
<tr>
<td>Token Ring</td>
<td>16 Mbps</td>
</tr>
<tr>
<td>Ethernet</td>
<td>10 Mbps</td>
</tr>
</tbody>
</table>

#### Wireless

<table>
<thead>
<tr>
<th>Technology</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>WiFi 802.11a</td>
<td>23 Mbps</td>
</tr>
<tr>
<td>WiFi 802.11g</td>
<td>19 Mbps</td>
</tr>
<tr>
<td>WiFi 802.11b</td>
<td>4.5 Mbps</td>
</tr>
</tbody>
</table>

### PAN technologies

<table>
<thead>
<tr>
<th>Technology</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>USB3</td>
<td>5 Gbps</td>
</tr>
<tr>
<td>UWB: Bluetooth3, Wireless USB, Wireless FireWire</td>
<td>480 Mbps</td>
</tr>
<tr>
<td>USB2</td>
<td>480 Mbps</td>
</tr>
<tr>
<td>USB1</td>
<td>12 Mbps</td>
</tr>
<tr>
<td>Bluetooth2</td>
<td>2 Mbps</td>
</tr>
<tr>
<td>Bluetooth1</td>
<td>723 Kbps</td>
</tr>
</tbody>
</table>

### Wide-area wireless data technologies

<table>
<thead>
<tr>
<th>Technology</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>WiMax 802.16</td>
<td>75 Mbps</td>
</tr>
<tr>
<td>WiBro (SKorea, mobile WiMax) 802.16</td>
<td>30-50 Mbps</td>
</tr>
<tr>
<td>CDMA 1xEVDO rev. B</td>
<td>5 Mbps</td>
</tr>
<tr>
<td>HSDPA</td>
<td>3.6 Mbps</td>
</tr>
<tr>
<td>CDMA 1xEVDO rev. A</td>
<td>3 Mbps</td>
</tr>
<tr>
<td>EDGE</td>
<td>236 Kbps</td>
</tr>
<tr>
<td>CDMA1x</td>
<td>163 Kbps</td>
</tr>
<tr>
<td>GPRS</td>
<td>80 Kbps</td>
</tr>
</tbody>
</table>

### Cable bandwidth

<table>
<thead>
<tr>
<th>Technology</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat-5 cable</td>
<td>100 Mbps</td>
</tr>
<tr>
<td>Cat-4 cable</td>
<td>20 Mbps</td>
</tr>
<tr>
<td>Cat-3 cable</td>
<td>10 Mbps</td>
</tr>
<tr>
<td>Level 1 cable (min)</td>
<td>20 Kbps</td>
</tr>
</tbody>
</table>