Peer-to-Peer File Sharing

CS 360 Internet Programming

Daniel Zappala

Brigham Young University
Computer Science Department

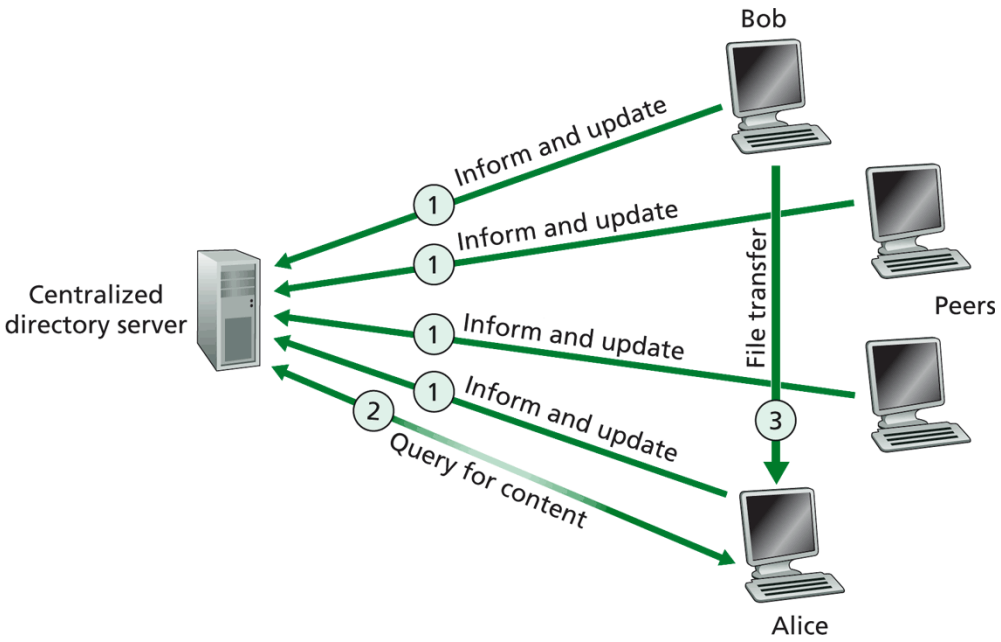
Overview

- peer-to-peer
  - peers exchange data with each other
  - peers act as both clients and servers
- file sharing
  - exchanging files between users
  - usually requires searching to find an available peer who has the file you want, then downloading that file

Peer-to-Peer Problems

- *bootstrapping*: finding peers to connect to
- *peer discovery*: finding other peers in the system
- *content location*: finding a peer with the desired content
- *content delivery*: downloading the content from a selected peer or peers

Napster



- bootstrapping, peer discovery: centralized server
- content location
  - tell server your IP address, filenames
  - send query to server, it returns list of peers with matching files
- content delivery: download file directly from a single peer

Copyright Law

- copyright: owner has exclusive rights to reproduce, adapt, publicly distribute, perofrm, and display their work
  - direct infringement: copying part or all of a copyrighted work without authorization
  - vicariaous liability: operator has (1) the right and ability to control users and (2) a direct financial benefit from allowing their acts of piracy.
  - contributory infringement: requires (1) knowledge of the infringing activity and (2) a material contribution – actual assistance or inducement – to the alleged piracy.

Fair Use

- use or copying of all or a portion of a copyrighted work without permission of the owner, e.g. for criticism, comment, news reporting, teaching, scholarship, or research
- courts consider:
  - purpose and character of use (commercial vs non profit)
  - nature of work
  - amount and substantiality of portion used (including size and quality)
  - the effect of use on market for or value of copyrighted work

Napster in Court

- Napster claims they are not infringing copyright because they are not storing any songs
- shutdown by court injunction because case against them was likely to succeed
  - Napster users likely guilty of *direct* copyright infringement - copying of a work by another
  - Napster likely to be guilty of *contributory* infringement because they learned of infringement and failed to purge the materials from its system
  - Napster likely to be guilty of *vicarious* infringement because they supervised or controlled the party engaging in infringing activity and had a financial interest in the activities
- see Wikipedia for background information

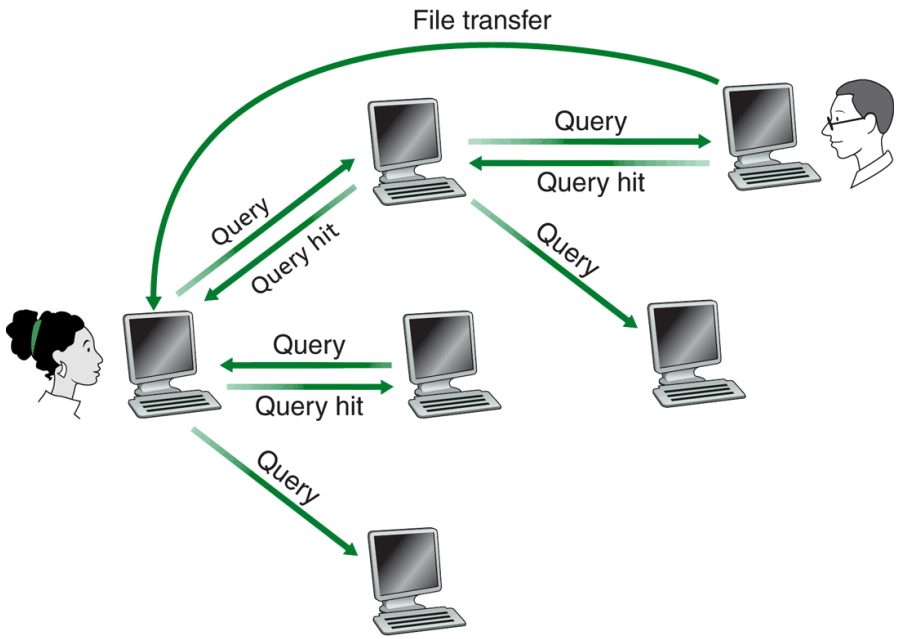
Promotional Power of Free Music

- record companies have claimed that free downloads supress sales
- some proof of the opposite effect
  - April 2000: tracks from Radiohead's *Kid A* album on Napster three months before CD release
  - millions of downloads by the time the record is released
  - number one spot on the charts in debut week, had never been in the top 20 before
  - beat many other heavily marketed artists
- **this example doesn't excuse piracy, but it does indicate that file sharing can provide a marketing opportunity for new bands**

Gnutella – version 0.4

- demise of Napster: copyright infringment (centralized solution makes it easy to find someone to sue)
- Gnutella: completely decentralized solution
- bootstrapping
  - first time: connect to a peer you heard about outside the system
  - for example, in a chat room
  - keep a cache of all peers discovered and use for bootstrapping next time
- peer discovery
  - try to always be connected to a fixed number of peers (TCP)
  - send a Ping message to existing neighbors, which is flooded to their neighbors
  - other peers respond to Ping with one or more Pong messages, containing IP address, port number, number of files sharing, number of KB sharing

Gnutella – version 0.4

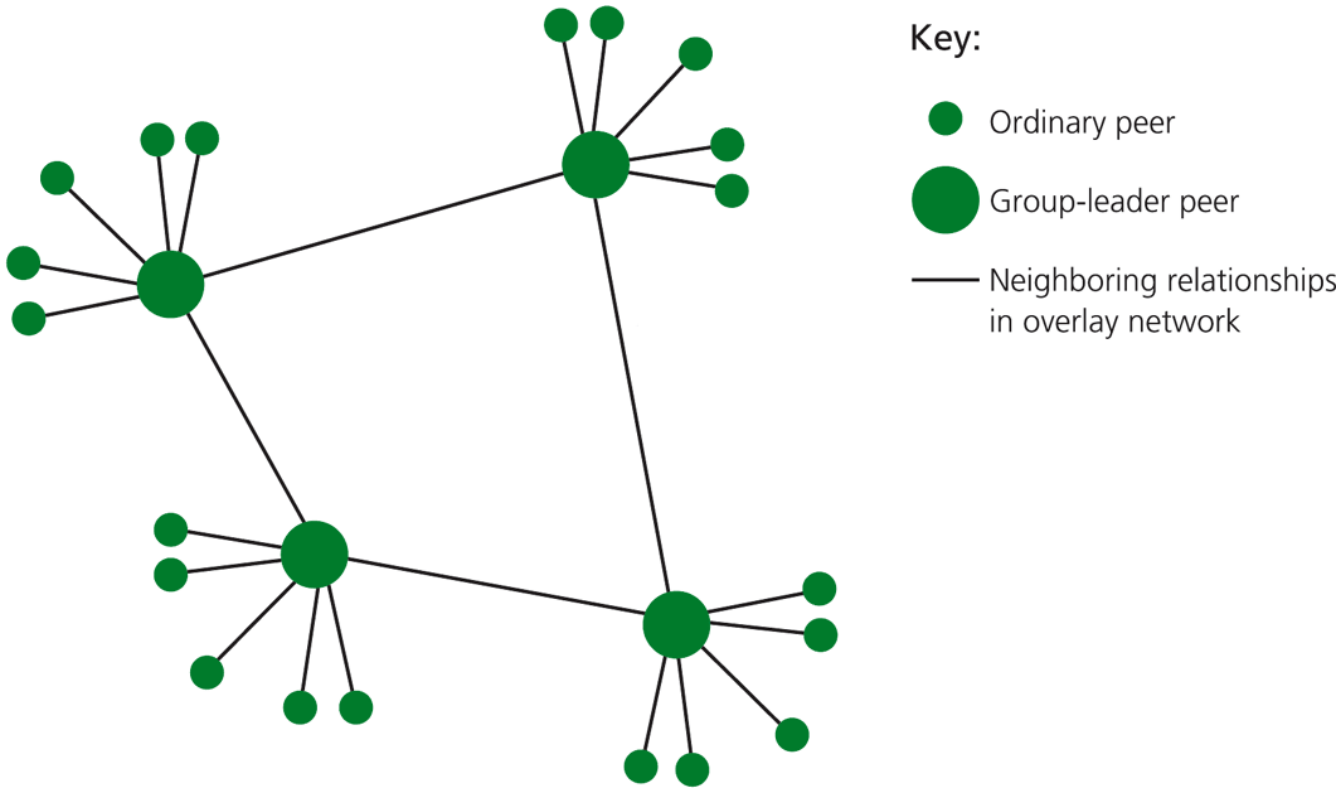


- content location
  - send a query to your neighbors
  - neighbors flood query, limited by a TTL
  - query includes minimum speed in kb/s for responding peers, search criteria (a string)
  - peers may respond with a query hit, listing IP address, port, number of hits, speed, result set (file name, size)
  - query hits sent along reverse path

Gnutella – version 0.4

- content delivery
  - direct download from a peer
  - if peer is behind a firewall, ask it to open a connection to you
  - if you are both behind a firewall, you are out of luck
- problems with Gnutella
  - no explicit rate limit on ping frequency or query frequency - quickly leads to overload
  - slow peers can hinder faster peers

Use Hierarchy To Scale



- use peers with high bandwidth as group leaders
- low-bandwidth peers connect to a group leader
- group leaders cache pointers to content at children
- queries sent among only the group leaders

Gnutella Ultrapeer Election

- self-nominate if eligible
  - not firewalled
  - suitable operating system
  - sufficient bandwidth
  - sufficient uptime
- leaf connects to an ultrapeer: drop all 0.4 connections
- leaf connects to an existing leaf: leaf re-routes to ultrapeer
- leaf connects to a leaf that cannot find an ultrapeer: allow 0.4 connection
- ultrapeer connects to an ultrapeer
  - if both have leaf nodes: make connection
  - if one has no leaf connections and too many ultrapeers: tell it to be a leaf
  - leaves should connect to at most 3 ultrapeers, really really should not allow more than 10 ultrapeeres

Gnutella Ultrapeer Query Routing

- leaves hash file names into a hash table with a "present" flag
  - also hash shortened names (to remove plurals)
- leaves send a compressed hash table to the ultrapeer
- ultrapeers route queries using version 4 flooding, but only send queries to client nodes if hash table hits

Gnutella – version 0.6

- ultrapeers
- pong caching
- network crawling
- query hit includes estimate of upload speed
- may deny downloads if too busy
- bye message for closing connections, debugging
- arbitrary protocol extensions
  - parallel download
  - tree hash exchange - verify blocks of a file
  - HUGE - persistent, location independent names (URNs)
  - LAN multicast