Peer to Peer I
Some of the peer-to-peer systems take distribution to an extreme, often for reasons other than technical.
What is different in distributed systems?

- Concurrency / Parallelism
- If done right, more resilient (e.g., Internet)
- But…
  - Partial and independent failures
  - Different parts of the system can be compromised
  - Different notions of time, order
- Our job is to design systems and algorithms that work despite all this

A distributed system is one in which the failure of a computer you didn’t even know existed can render your own computer unusable. Leslie Lamport

Peer to peer systems exacerbate all of these characteristics
Lamport quote: http://lamport.azurewebsites.net/pubs/distributed-system.txt
In the late 90’s two trends created the conditions for a new killer application for the Internet: music sharing. The two trends were the availability of broadband Internet, and the advent of good-quality audio compression (mp3, 1:12).
Peer-to-Peer Systems

- How did it start?
  - A killer application: file distribution
  - Free music over the Internet! (not exactly legal...)

- Key idea: share storage, content, and bandwidth of individual users
  - Lots of them

- Big challenge: coordinate all of these users
  - In a scalable way (not NxN!)
  - With changing population (aka churn)
  - With no central administration
  - With no trust
  - With large heterogeneity (content, storage, bandwidth, ...)

---

CS 138
II–5

Copyright © 2015 Thomas W. Duquette, Rodrigo Fonseca
All rights reserved.
3 Key Requirements

- P2P Systems do three things:
  - Help users determine what they want
    - Some form of search
    - P2P version of Google
  - Locate that content
    - Which node(s) hold the content?
    - P2P version of DNS (map name to location)
  - Download the content
    - Should be efficient
    - P2P form of Akamai
The Napster file sharing service features a central service at which providers of files register their locations and seekers of files find file locations.
A central server is clearly a bottleneck. But we can replicate it locally, as we will see. The main problem, as was discovered by the providers of the original Napster, its presence makes it easy for legal action to be taken against the service.
So a radically different approach appeared: a totally decentralized architecture.
Some Details

- Participants interconnect via overlay network
- To send a query:
  - send request to each directly connected node
  - proceed for some maximum number of hops
    - node having desired file sends back its identity
      • over reverse query route in original Gnutella
      • direct via UDP in later Gnutella
    - querier chooses a source (if necessary)
    - sends it a push request
      • transfer via HTTP

See the Wikipedia article (http://en.wikipedia.org/wiki/Gnutella) for a few more details.
More Details

- Joining the overlay network:
  - obtain addresses of some number of network nodes
    - wired into code
    - check web site
    - etc.
  - contact them; they produce address of other nodes
  - connect to \( n \) of them
  - keep others cached for later use
Problems

- Flaky network connections
- Flaky computers
- Flaky users
This architecture later led to Kazaa, and to Skype!
Lessons and Limitations

- Client-server simple and effective
  - But not always feasible
- Things that flood-based systems do well
  - Decentralization of visibility and liability
  - Finding popular stuff
  - Fancy local queries
- Things that flood-based systems do poorly
  - Scale (exponential increase in traffic vs hops)
  - Finding unpopular stuff
  - Fancy distributed queries
  - Vulnerabilities: data poisoning, tracking, etc.
  - Guarantees about anything (answer quality, privacy, etc.)
Second generation P2P

- Structured P2P systems, mostly academic efforts
- Goal: solve the scalable decentralized location problem
Distributed Hash Tables

hash("metal heart")
Straw man: modulo hashing

- Say you have N servers
- Map requests to servers as follows:
  - Number servers 0 to N-1
  - Compute hash of content: \( h = \text{hash (name)} \)
  - Redirect client to server \( #p = h \mod N \)
- Keep track of load in each proxy
  - If load on proxy \( #p \) is too high, try again with a different hash function (or “salt”)
- Problem: most caches will be useless if you add or remove proxies, change value of N
Consistent Hashing [Karger et al., 99]

- Servers and objects mapped to points on a circle using hash
- An object is assigned to its successor server
- Minimizes data movement on change!
  - Only $O(1/N)$ objects moved on server leave/join
  - Which ones?
A paper explaining Chord can be found at http://pdos.csail.mit.edu/papers/chord:sigcomm01/chord_sigcomm.pdf. The hash function employed is SHA-1. The bounds on the number of messages are “with high probability.”
The range of the hash function is organized as a circle. The red circles represent nodes (computers) whose hashed IP addresses are 0, 1, and 3. To simplify the discussion, we’ll ignore the hash function and refer to 0, 1, and 3 as being the nodes themselves rather than their hashed IP addresses. Similarly, we’ll refer to keys 0, 1, … 7 rather than saying that we have keys whose values hash to [0, 7]. Given this simplification, if \( i \) is a key, then \( \text{successor}(i) \) is the node where the key (and associated value) is stored. Things are organized so that key \( i \) is assigned to the lowest numbered node greater than or equal to \( i \) (modulo \( 2^m \)). Thus \( \text{successor}(i) \) is the number of that node.

If we store with each node \( i \) the address of \( \text{successor}(i) \), then, starting from any node, we can find the node containing any particular key (or definitively say the key is not present). Of course, doing this requires \( O(n) \) steps.
Search requires a number of messages that is linear in the number of nodes — not good.
Note that the numbering of finger-table entries starts with 1 (not 0)!
With the addition of the finger table, search requires $\log(N)$ messages, where $N$ is the number of nodes.
With the addition of the finger table, search requires $\log(N)$ messages, where $N$ is the number of nodes.
1. While not in id’s predecessor
2. Find last finger $f < id$
3. Recurse
In this pseudo code (taken from the aforementioned paper), \( m.\text{foo}(x) \) means to place a remote procedure call to node \( m \), executing procedure \( \text{foo} \) with argument \( x \). \( m.x \) means to place a remote procedure call to node \( m \), retrieving the value of variable \( x \).

The while loop in \( \text{find_predecessor} \) continues until \( n1 \) is the highest-numbered node less than \( id \) (modulo \( 2^m \)), which will be the case when \( id \) is between \( n1 \) and the node that comes after it (\( n1.\text{successor} \)).

The for loop in \( \text{closest_preceding_finger} \) finds the highest-numbered row in the finger table that refers to a node that comes before \( id \).
Here are the finger tables for our example. Each row represents the portion of the key space covered by the row ("finger").

<table>
<thead>
<tr>
<th>j</th>
<th>start</th>
<th>interval</th>
<th>node</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>[1, 2]</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>[2, 4]</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>[4, 0]</td>
<td>0</td>
</tr>
<tr>
<td>predecessor</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>j</th>
<th>start</th>
<th>interval</th>
<th>node</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>[2, 3]</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>[3, 5]</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>[5, 1]</td>
<td>0</td>
</tr>
<tr>
<td>predecessor</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>j</th>
<th>start</th>
<th>interval</th>
<th>node</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>[4, 5]</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>[5, 7]</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>[7, 3]</td>
<td>0</td>
</tr>
<tr>
<td>predecessor</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that the node listed in the first row of each table is the next node in the ring. This is somewhat confusingly called the “successor node” (if x is a hashed key, then successor(x) might be x if node x exists; but successor(x) if x is a node is always the next node in the ring). It’s convenient to also list the predecessor node for each node; it’s given at the end of each table.
Here we include in the finger tables the objects stored at each node.
Adding a node requires both adjusting all the finger tables to accommodate the new node and moving the objects that should be stored at that node.
First we set up the finger table of the new node.
Next we modify the finger tables of previously existing nodes to accommodate the new node.
Finally we redistribute the objects.
Invariants

- Each node’s successor link is correct
- For every key k, successor(k) is responsible for k
For the moment, we ignore finger tables. This code provides the minimum necessary functionality so that the first invariant is preserved. A real implementation could be more aggressive. To make sense of this code, see the next few slides.
Focusing just on the successor and predecessor links, we follow what happens when node 6 is added to the ring.
Node 6 calls join to add itself. At this point, no other node knows of its existence.
Node 6 now calls stabilize. Stabilize itself does nothing, but it calls notify on node 0 (6’s successor). Node 0 sets its predecessor to be 6.
Node 3 now calls stabilize and discovers that its successor’s predecessor is not itself. So it sets its successor to 6 (its successor’s predecessor) and calls notify on node 6. Node 6 now sets its predecessor to be 3 — it’s now fully linked in.
Transferring Objects

- When?
  - not until new node is fully linked in
  - could be a race between a search and the transfer
- What to do?
  - if search fails, search again after a delay
Finger Tables?

- If finger tables aren’t updated, is correctness affected?
This slide shows the effect of adding nodes 6 and 4 in our earlier example by merely executing the stabilization code.

Note that the finger table is used to find the predecessor of the object’s successor. Thus searches initiated at nodes 0 and 1 for object 5 will identify node 3 as the predecessor. It’s then necessary to follow successor links until 5 is located at its successor, node 6.
Updating Finger Tables

// this is run periodically
n.fix_next_finger()
  // i is initialized to 1 outside of the function
  finger[i].node = find_successor(finger[i].start)
  i++
  if i > m - 1
    i = 1
What to Do?

- Each node keeps list of $r$ nearest successors
  - if one does not respond, switch to next
- Also replicate data at the $r$ successors
We didn’t cover

- Detailed failed recovery
- In 2012, using a formal model of Chord in Alloy, Pamela Zave showed that Chord is not correct!
  - E.g., multiple simultaneous joins can result in wrong order
  - Node leaving and then rejoining with the same id can lead to node pointing at itself
  - Has a version of the spec she claims correct
- Very subtle bugs, took over 10 years to find, over 2000 citations, “Test of Time” award
- If you are interested, take Logic for Systems, 1950-Y!

This can be seen here: http://www2.research.att.com/~pamela/chord.html
Next Class

- Tapestry, another DHT
- Chord assignment due Feb 14th, Tuesday!

- Make sure you have done:
  - Collaboration policy
  - Piazza
  - Github