CS5412: BIMODAL MULTICAST ASTROLABE

Lecture XIX
Gossip 201

- Recall from early in the semester that gossip spreads in $\log$ (system size) time.
- But is this actually “fast”? 
Gossip in distributed systems

- Log(N) can be a very big number!
  - With N=100,000, log(N) would be 12
  - So with one gossip round per five seconds, information needs one minute to spread in a large system!

- Some gossip protocols combine pure gossip with an accelerator
  - A good way to get the word out quickly
Bimodal Multicast

- To send a message, this protocol uses IP multicast
- We just transmit it without delay and we don’t expect any form of responses
  - Not reliable, no acks
  - No flow control (this can be an issue)
  - In data centers that lack IP multicast, can simulate by sending UDP packets 1:1 without acks
What’s the cost of an IP multicast?

- In principle, each Bimodal Multicast packet traverses the relevant data center links and routers just once per message.

- So this is extremely cheap... but how do we deal with systems that didn’t receive the multicast?
Making Bimodal Multicast reliable

- We can use gossip!

- Every node tracks the membership of the target group (using gossip, just like with Kelips, the DHT we studied early in the semester)
  - Bootstrap by learning “some node addresses” from some kind of a server or web page
  - But then exchange of gossip used to improve accuracy
Now, layer in a gossip mechanism that gossips about multicasts each node knows about.

Rather than sending the multicasts themselves, the gossip messages just talk about “digests”, which are lists.

- Node A might send node B
  - I have messages 1-18 from sender X
  - I have message 11 from sender Y
  - I have messages 14, 16 and 22-71 from sender Z

- Compactly represented...

- This is a form of “push” gossip
On receiving such a gossip message, the recipient checks to see which messages it has that the gossip sender lacks, and vice versa.

Then it responds:
- I have copies of messages $M$, $M'$, and $M''$ that you seem to lack
- I would like a copy of messages $N$, $N'$, and $N''$ please

An exchange of the actual messages follows.
Bimodal Multicast resends using IP multicast if there is “evidence” that a few nodes may be missing the same thing.
- E.g. if two nodes ask for the same retransmission
- Or if a retransmission shows up from a very remote node (IP multicast doesn’t always work in WANs)

It also prioritizes recent messages over old ones.

Reliability has a “bimodal” probability curve: either nobody gets a message or nearly everyone does.
In this variation on Bimodal Multicast instead of gossiping with every node in a system, we modify the Bimodal Multicast protocol

- It maintains a “peer overlay”: each member only gossips with a smaller set of peers picked to be reachable with low round-trip times, plus a second small set of remote peers picked to ensure that the graph is very highly connected and has a small diameter

- Called a “small worlds” structure by Jon Kleinberg

- Lpbcast is often faster, but equally reliable!
Speculation... about speed

- When we combine IP multicast with gossip we try to match the tool we’re using with the need

- Try to get the messages through fast... but if loss occurs, try to have a very predictable recovery cost
  - Gossip has a totally predictable worst-case load
  - This is appealing at large scales

- How can we generalize this concept?
A thought question

What’s the best way to

- Count the number of nodes in a system?
- Compute the average load, or find the most loaded nodes, or least loaded nodes?

Options to consider

- Pure gossip solution
- Construct an overlay tree (via “flooding”, like in our consistent snapshot algorithm), then count nodes in the tree, or pull the answer from the leaves to the root…
... and the answer is

- Gossip isn’t very good for some of these tasks!
  - There are gossip solutions for counting nodes, but they give approximate answers and run slowly
  - Tricky to compute something like an average because of “re-counting” effect, (best algorithm: Kempe et al)

- On the other hand, gossip works well for finding the c most loaded or least loaded nodes (constant c)

- Gossip solutions will usually run in time O(log N) and generally give probabilistic solutions
Recall how flooding works

Basically: we construct a tree by pushing data towards the leaves and linking a node to its parent when that node first learns of the flood.

Can do this with a fixed topology or in a gossip style by picking random next hops.

Labels: distance of the node from the root.
Once we have a spanning tree

- To count the nodes, just have leaves report 1 to their parents and inner nodes count the values from their children
- To compute an average, have the leaves report their value and the parent compute the sum, then divide by the count of nodes
- To find the least or most loaded node, inner nodes compute a min or max...

Tree should have roughly $\log(N)$ depth, but once we build it, we can reuse it for a while
Not all logs are identical!

- When we say that a gossip protocol needs time $\log(N)$ to run, we mean $\log(N)$ rounds.
  - And a gossip protocol usually sends one message every five seconds or so, hence with 100,000 nodes, 60 secs.

- But our spanning tree protocol is constructed using a flooding algorithm that runs in a hurry.
  - $\log(N)$ depth, but each “hop” takes perhaps a millisecond.
  - So with 100,000 nodes we have our tree in 12 ms and answers in 24ms!
Insight?

- Gossip has time complexity $O(\log N)$ but the “constant” can be rather big (5000 times larger in our example)

- Spanning tree had same time complexity but a tiny constant in front

- But network load for spanning tree was much higher
  - In the last step, we may have reached roughly half the nodes in the system
  - So 50,000 messages were sent all at the same time!
Gossip vs “Urgent”? 

- With gossip, we have a slow but steady story
  - We know the speed and the cost, and both are low
  - A constant, low-key, background cost
  - And gossip is also very robust

- Urgent protocols (like our flooding protocol, or 2PC, or reliable virtually synchronous multicast)
  - Are way faster
  - But produce load spikes
  - And may be fragile, prone to broadcast storms, etc
Introducing hierarchy

- One issue with gossip is that the messages fill up
  - With constant sized messages...
  - ... and constant rate of communication
  - ... we’ll inevitably reach the limit!

- Can we introduce hierarchy into gossip systems?
Intended as help for applications adrift in a sea of information

Structure emerges from a randomized gossip protocol

This approach is robust and scalable even under stress that cripples traditional systems

Developed at RNS, Cornell

By Robbert van Renesse, with many others helping...

Technology was adopted at Amazon.com (but they build their own solutions rather than using it in this form)
Astrolabe is a flexible monitoring overlay

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>2271</td>
<td>1.8</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>1971</td>
<td>1.5</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>2004</td>
<td>4.5</td>
<td>1</td>
<td>0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

Periodically, pull data from monitored systems

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>2003</td>
<td>.67</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>1976</td>
<td>2.7</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>2231</td>
<td>1.7</td>
<td>1</td>
<td>1</td>
<td>6.0</td>
</tr>
</tbody>
</table>
Astrolabe in a single domain

- Each node owns a single tuple, like the management information base (MIB)
- Nodes discover one-another through a simple broadcast scheme ("anyone out there?") and gossip about membership
  - Nodes also keep replicas of one-another’s rows
  - Periodically (uniformly at random) merge your state with some else...
State Merge: Core of Astrolabe epidemic

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>2011</td>
<td>2.0</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>1971</td>
<td>1.5</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>2004</td>
<td>4.5</td>
<td>1</td>
<td>0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>2003</td>
<td>.67</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>1976</td>
<td>2.7</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>2201</td>
<td>3.5</td>
<td>1</td>
<td>1</td>
<td>6.0</td>
</tr>
</tbody>
</table>
State Merge: Core of Astrolabe epidemic

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>2011</td>
<td>2.0</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>1971</td>
<td>1.5</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>2004</td>
<td>4.5</td>
<td>1</td>
<td>0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>2003</td>
<td>.67</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>1976</td>
<td>2.7</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>2201</td>
<td>3.5</td>
<td>1</td>
<td>1</td>
<td>6.0</td>
</tr>
</tbody>
</table>

swift.cs.cornell.edu

cardinal.cs.cornell.edu

CS5412 Spring 2016
State Merge: Core of Astrolabe epidemic

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>2011</td>
<td>2.0</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>1971</td>
<td>1.5</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>2201</td>
<td>3.5</td>
<td>1</td>
<td>0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

CS5412 Spring 2016
Observations

- Merge protocol has constant cost
  - One message sent, received (on avg) per unit time.
  - The data changes slowly, so no need to run it quickly – we usually run it every five seconds or so
  - Information spreads in $O(\log N)$ time
- But this assumes bounded region size
  - In Astrolabe, we limit them to 50-100 rows
A big system could have many regions

- Looks like a pile of spreadsheets
- A node only replicates data from its neighbors within its own region
Scaling up… and up…

- With a stack of domains, we don’t want every system to “see” every domain
  - Cost would be huge
- So instead, we’ll see a summary
Astrolabe builds a hierarchy using a P2P protocol that “assembles the puzzle” without any servers

Dynamically changing query output is visible system-wide

SQL query “summarizes” data

<table>
<thead>
<tr>
<th>Name</th>
<th>Avg Load</th>
<th>WL contact</th>
<th>SMTP contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
<td>2.2</td>
<td>123.45.61.3</td>
<td>123.45.61.17</td>
</tr>
<tr>
<td>Nj</td>
<td>1.6</td>
<td>127.16.77.6</td>
<td>127.16.77.11</td>
</tr>
<tr>
<td>Paris</td>
<td>2.7</td>
<td>14.66.71.8</td>
<td>14.66.71.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>1.7</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>2.1</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>3.9</td>
<td>1</td>
<td>0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>gazelle</td>
<td>4.1</td>
<td>0</td>
<td>0</td>
<td>4.5</td>
</tr>
<tr>
<td>zebra</td>
<td>0.9</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>gnu</td>
<td>2.2</td>
<td>1</td>
<td>0</td>
<td>6.2</td>
</tr>
</tbody>
</table>
Large scale: “fake” regions

- These are
  - Computed by queries that summarize a whole region as a single row
  - Gossiped in a read-only manner within a leaf region

- But who runs the gossip?
  - Each region elects “k” members to run gossip at the next level up.
  - Can play with selection criteria and “k”
Hierarchy is virtual... data is replicated

Yellow leaf node “sees” its neighbors and the domains on the path to the root.

Falcon runs level 2 epidemic because it has lowest load.

Gnu runs level 2 epidemic because it has lowest load.

<table>
<thead>
<tr>
<th>Name</th>
<th>Load</th>
<th>Avg Load</th>
<th>WL contact</th>
<th>SMTP contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
<td>2.6</td>
<td>123.45.61.3</td>
<td>123.45.61.17</td>
<td></td>
</tr>
<tr>
<td>NJ</td>
<td>1.8</td>
<td>127.16.77.6</td>
<td>127.16.77.11</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Avg</th>
<th>WL</th>
<th>SMTP</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>gazelle</td>
<td>1.7</td>
<td>0</td>
<td>0</td>
<td>4.5</td>
</tr>
<tr>
<td>zebra</td>
<td>3.2</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>gnu</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>6.2</td>
</tr>
</tbody>
</table>

San Francisco

New Jersey

CS5412 Spring 2016
Hierarchy is virtual... data is replicated.

<table>
<thead>
<tr>
<th>Name</th>
<th>Avg Load</th>
<th>WL contact</th>
<th>SMTP contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
<td>2.6</td>
<td>123.45.61.3</td>
<td>123.45.61.17</td>
</tr>
<tr>
<td>Nj</td>
<td>1.8</td>
<td>127.16.77.6</td>
<td>127.16.77.11</td>
</tr>
<tr>
<td>Paris</td>
<td>3.1</td>
<td>14.66.71.8</td>
<td>14.66.71.12</td>
</tr>
</tbody>
</table>

Green node sees different leaf domain but has a consistent view of the inner domain.

San Francisco:

<table>
<thead>
<tr>
<th>Name</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>2.0</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>1.5</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>4.5</td>
<td>1</td>
<td>0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

New Jersey:

<table>
<thead>
<tr>
<th>Name</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>gazelle</td>
<td>1.7</td>
<td>0</td>
<td>0</td>
<td>4.5</td>
</tr>
<tr>
<td>zebra</td>
<td>3.2</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>gnu</td>
<td>.5</td>
<td>1</td>
<td>0</td>
<td>6.2</td>
</tr>
</tbody>
</table>

CS5412 Spring 2016
Worst case load?

- A small number of nodes end up participating in $O(\log_{\text{fanout}} N)$ epidemics
  - Here the fanout is something like 50
  - In each epidemic, a message is sent and received roughly every 5 seconds
- We limit message size so even during periods of turbulence, no message can become huge.
Who uses Astrolabe?

- Amazon doesn’t use Astrolabe in this identical form, but they built gossip-based monitoring systems based on the same ideas.
- They deploy these in S3 and EC2: throughout their big data centers!
  - For them, Astrolabe-like mechanisms track overall state of their system to diagnose performance issues
  - They also automate reaction to temporary overloads
Example of overload handling

- Some service S is getting slow...
  - Astrolabe triggers a “system wide warning”

- Everyone sees the picture
  - “Oops, S is getting overloaded and slow!”
  - So everyone tries to reduce their frequency of requests against service S

- What about overload in Astrolabe itself?
  - Could everyone do a fair share of inner aggregation?
Idea that one company had

- Start with Astrolabe approach

- But instead of electing nodes to play inner roles, just assign them roles, left to right

- N-1 inner nodes, hence N-1 nodes play 2 aggregation roles and one lucky node just has one role

- What impact will this have on Astrolabe?
World’s worst aggregation tree!

An event e occurs at H

G gossips with H and learns e

P learns O(N) time units later!
What went wrong?

- In this horrendous tree, each node has equal “work to do” but the information-space diameter is larger!
- Astrolabe benefits from “instant” knowledge because the epidemic at each level is run by someone elected from the level below
Insight: Two kinds of shape

- We’ve focused on the aggregation tree
- But in fact should also think about the information flow tree
Information space perspective

- **Bad aggregation graph:** diameter $O(n)$

- **Astrolabe version:** diameter $O(\log(n))$
Where will this approach go next?

- Cornell research: Shared State Table (SST)

- Idea is to use RDMA to implement a super-fast shared state table for use in cloud data centers

- Tradeoff: simplified model but WAY faster
SST details

- Starts like Astrolabe with a table for a region
  - But no hierarchy, SST is like a “one level” Astrolabe
  - And the implementation doesn’t use gossip....

- To understand SST, you need to know about RDMA
  - This is a new kind of hardware
  - Allows DMA transfers over the optical network from sender memory to receiver memory, at full network data rates which can be 10Gbps or even 100Gbps.

<table>
<thead>
<tr>
<th>Name</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>gazelle</td>
<td>1.7</td>
<td>0</td>
<td>0</td>
<td>4.5</td>
</tr>
<tr>
<td>zebra</td>
<td>3.2</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>gnu</td>
<td>.5</td>
<td>1</td>
<td>0</td>
<td>6.2</td>
</tr>
</tbody>
</table>
RDMA Concept

- RDMA is a new standard for Reliable DMA

Bits moved by the optical network directly from sender memory to receiver memory
RDMA acronyms: Enliven a party...

- **RDMA**: Short for Remote Direct Memory Access. Offers two modes:
  - **Block transfers** (can be very large): binary from sender to destination; sender specifies source and destination provides a receive buffer
  - **One-sided read**: Receiver allows sender to do a read from memory without any software action on receiver side. **One-sided write**: same idea, but allows writes

- **Three ways to obtain RDMA**:
  - **RDMA on Infiniband** network hardware: popular in HPC
  - **RDMA on fast Ethernet**: also called RDMA/ROCE. Very new and we need experience to know how well this will work. Issue centers on a new approach to flow control.
  - **RDMA in software**: SoftROCE. A fake RDMA for running RDMA code on old machines lacking RDMA hardware. Tunnels on TCP

- **IB Verbs**: this is the name for the software library used to access RDMA
SST uses RDMA to replicate rows

- With SST, when a node updates its row, RDMA is used to copy the row to other nodes.

- Then the compute model offers predicates on the table, aggregation, and upcalls to event handlers:
  - As many as 1M or 2M, per node, per second.
  - This is so much faster than Astrolabe that comparison is kind of meaningless... maybe 100,000x faster...
SST: Like Astrolabe, but one region, fast updates

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Load</th>
<th>Weblogic?</th>
<th>SMTP?</th>
<th>Word Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>swift</td>
<td>2011</td>
<td>2.0</td>
<td>0</td>
<td>1</td>
<td>6.2</td>
</tr>
<tr>
<td>falcon</td>
<td>1971</td>
<td>1.5</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>cardinal</td>
<td>2201</td>
<td>3.5</td>
<td>1</td>
<td>1</td>
<td>6.0</td>
</tr>
</tbody>
</table>

swift.cs.cornell.edu

falcon.cs.cornell.edu

Cardinal.cs.cornell.edu

CS5412 Spring 2016
Basically, each program registers code like this:

```c
if (some-predicate-becomes true)
{
    do this “upcall” logic
}
```

SST evaluates predicates and issues upcalls in parallel, on all machines, achieving incredible speed.

After machine A changes some value, a predicate on machine B might fire in 1us or less delay…
## Use case comparison

<table>
<thead>
<tr>
<th><strong>Astrolabe</strong></th>
<th><strong>SST</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>□ Kind of slow</td>
<td>□ Crazy fast</td>
</tr>
<tr>
<td>□ Used at large scale to monitor big data centers, for “data mining”</td>
<td>□ Used mostly at rack scale: maybe 64 or 128 nodes at a time</td>
</tr>
<tr>
<td>□ Uses hierarchical structure to scale, gossip as its communication protocol</td>
<td>□ Has a flat structure, and uses RDMA to transfer rows directly at optical network speeds of 10 or even 100Gbps</td>
</tr>
</tbody>
</table>
Summary

- First we saw a way of using Gossip in a reliable multicast (although the reliability is probabilistic)
  - Astrolabe uses Gossip for aggregation
  - Pure gossip isn’t ideal for this… and competes poorly with flooding and other urgent protocols
  - But Astrolabe introduces hierarchy and is an interesting option that gets used in at least one real cloud platform

- SST uses a similar model but
  - Implemented with one-sided RDMA reads, writes
  - Less scalable, but runs 100,000x faster…