



Distributed Coordination via ZooKeeper

Flavio Junqueira
Yahoo! Research, Barcelona

Hadoop in China 2011

A bit of history



June 2007: Early adopters: Message Broker, Crawler

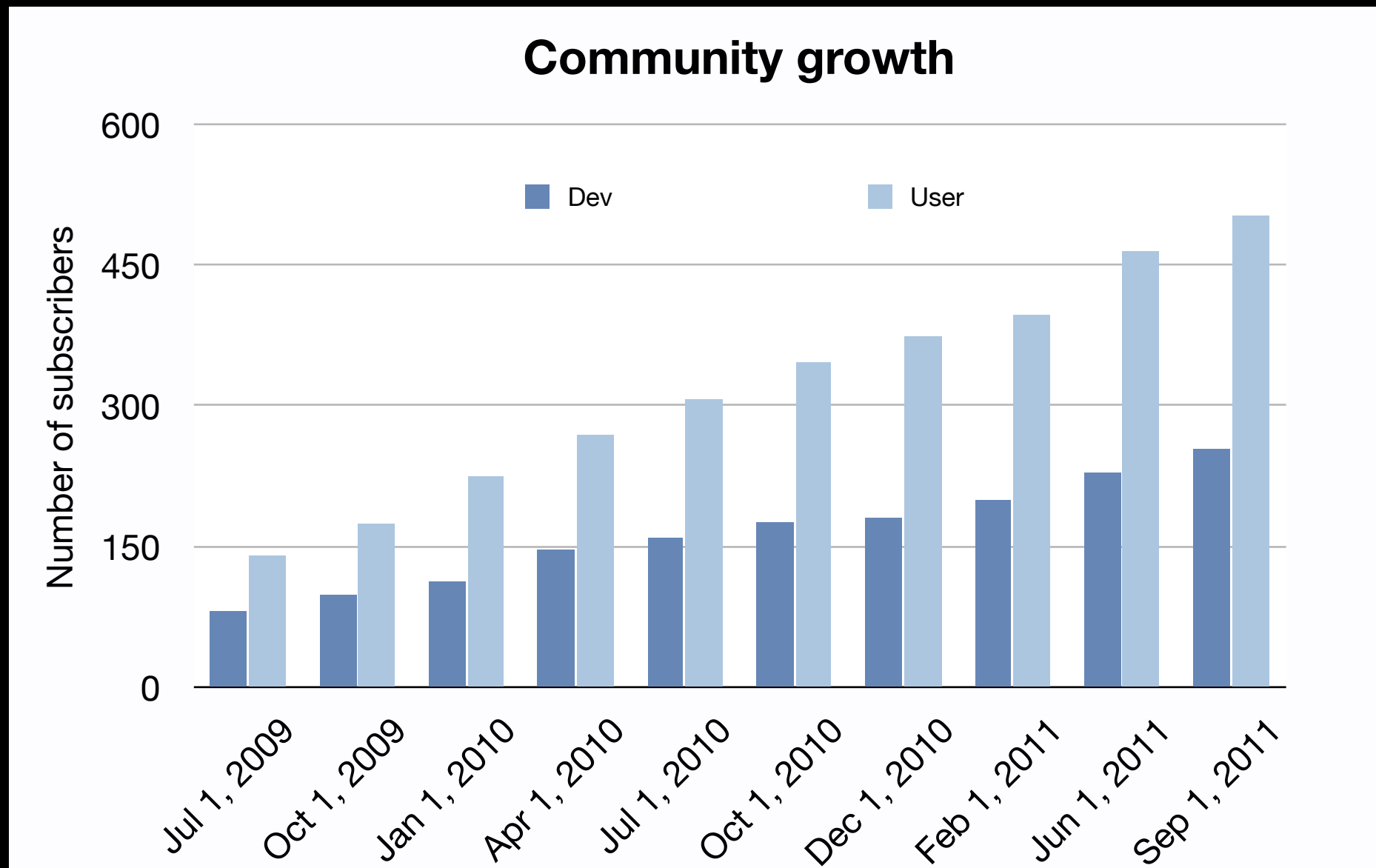
Oct 2007: Sourceforge

June 2008: Move to Apache, subproject of Hadoop

Nov 2010: Top level project



Apache community growth



Widely used



Widely used

{ 2010 08 26 }

Zookeeper experience

While working on Kafka, a distributed pub/sub system (more on that later) at LinkedIn, I need to use [Zookeeper](#) (ZK) to implement the load-balancing logic. I'd like to share my experience of using Zookeeper. First of all, for those of you who don't know, Zookeeper is an Apache project that implements a consensus service based on a variant of [Paxos](#) (it's similar to Google's [Chubby](#)). ZK has a very simple, file system like API. One can create a path, set the value of a path, read the value of a path, delete a path, and list the children of a path. ZK does a couple of more interesting things: (a) one can register a watcher on a path and get notified when the children of a path or the value of a path is changed, (b) a path can be created as ephemeral, which means that if the client that created the path is gone, the path is automatically removed by the ZK server. However, don't let the simple API fool you. One needs to understand a lot more than those APIs in order to use them properly. For me, this translates to weeks asking the ZK mailing list (which is pretty responsive) and our local ZK experts.

Jun Rao, LinkedIn



Widely used

{ 2010 08 26 }

Zookeeper experience

While working on Kafka, a distributed pub/sub system (more on that later) at LinkedIn, I need to use [Zookeeper](#) (ZK) to implement the load balancing logic. I'd like to share my experience of using Zookeeper. Zookeeper is based on a variant of [Paxos](#) (it's similar to Raft). It provides a path, delete a path, and list the children of a path or the value of a path. It also automatically removes the ZK session if it's not properly. For me, this translates to

Since Messages accepts data from many sources such as email and SMS, we decided to write an application server from scratch instead of using our generic Web infrastructure to handle all decision making for a user's messages. It interfaces with a large number of other services: we store attachments in [Haystack](#), wrote a user discovery service on top of [Apache ZooKeeper](#), and talk to other infrastructure services for email account verification, friend relationships, privacy decisions, and delivery decisions (for example, should a message be sent over chat or SMS). We spent a lot of time making sure each of these services are reliable, robust, and performant enough to handle a real-time messaging system.

Kannan Muthukkaruppan, Facebook



Widely used

{ 2010 08 26 }

Zookeeper experience

While working on Kafka, a distributed pub/sub system (more on that later) at LinkedIn, I need to use [Zookeeper](#) (ZK) to implement the load balancing logic. I'd like to share my experience of using Zookeeper. Zookeeper is based on a variant of [Paxos](#) (it's similar to the consensus algorithm). It provides a way to create a path, delete a path, and list the children of a path or the value of a path. Zookeeper also automatically removes the ZK session if it is not properly. For me, this translates to

Since Messages accepts data from many sources such as email and SMS, we decided to write an application server from scratch instead of using our generic Web infrastructure to handle all decision making for a user's messages. It interfaces with a large number of other services: we store attachments in [Haystack](#), wrote a user discovery service on top of [Apache](#)

- **ResourceManager** - The ResourceManager uses [Apache ZooKeeper](#) for fail-over. When the ResourceManager fails, a secondary can quickly recover via cluster state saved in ZooKeeper. The ResourceManager, on a fail-over, restarts all of the queued and running applications.

Arun Murty, NextGen Hadoop



Widely used

{ 2010 08 26 }

Zookeeper experience

While working on Kafka, a distributed pub/sub system (more on that later) at LinkedIn, I need to use [Zookeeper](#) (ZK) to implement the load balancing logic. I'd like to share my experience of using Zookeeper. Zookeeper is based on a variant of [Paxos](#) (it's similar to Raft). It provides a path, delete a path, and list the children of a path or the value of a path. It also provides a way to automatically removed by the ZK server. For me, this translates to

Since Messages accepts data from many sources such as email and SMS, we decided to write an application server from scratch instead of using our generic Web infrastructure to handle all decision making for a user's messages. It interfaces with a large number of other services: we store attachments in [Haystack](#), wrote a user discovery service on top of [Apache](#)

- **ResourceManager** - The ResourceManager uses [Apache ZooKeeper](#) for fail-over. When the ResourceManager fails, a secondary can quickly recover via cluster state saved in ZooKeeper. This is used by many applications.

The Apache Software Foundation has a neat tool for distributed lock services, called Zookeeper. Scalr based its distributed cron jobs off of it, so that users can setup scripts to be executed periodically, like cron jobs, without running the risk of multiple executions or failure to execute.

Sebastian Stadil, Scalr



Widely used

{ 2010 08 26 }

Zookeeper experience

While working on Kafka, a distributed pub/sub system (more on that later) at LinkedIn, I need to use [Zookeeper](#) (ZK) to implement the load balancing logic. I'd like to share my experience of using Zoo on a variant of [Paxos](#) (it's similar to path, delete a path, and list the children of a path or the value of a path automatically removed by the ZK service properly. For me, this translates to

Since Messages accepts data from many sources such as email and SMS, we decided to write an application server from scratch instead of using our generic Web infrastructure to handle all decision making for a user's messages. It interfaces with a large number of other services: we store attachments in [Haystack](#), wrote a user discovery service on top of [Apache](#)

- **ResourceManager** - The ResourceManager uses [Apache ZooKeeper](#) for fail-over. When the ResourceManager fails, a secondary can quickly recover via cluster state saved in ZooKeeper applications

The Apache Software Foundation has a neat tool for distributed lock services, called Zookeeper. Scalr based its distributed cron jobs off of it, so that users can setup scripts to be executed periodically, like cron jobs, without running the risk of multiple executions or failure to execute.

Operating System and Configuration

Digg runs on [Debian](#) stable based GNU/Linux servers which we configure with [Clusto](#), [Puppet](#) and using a configuration system over [Zookeeper](#).

Dave Beckett, Digg.com





Where we are coming from...

Yahoo! Portal

MY YAHOO! Web Images Video Local Shopping more

Web Search

Quicklinks My Front Page The Best of My Yahoo! NEW New Tab

Hi, Flavio Sign Out Tips Help

Add Content Change Appearance More Options My Yahoo! Blog: It's Y!ou

Personal Assistant Options

Mail Horoscope Stocks

Weather 57°F Lottery Sports

Message Center Options

Weather Options

Compact Classic Full

57°F Mostly Cloudy

Location	Today	Tomorrow	Monday
Philadelphia, PA* Mostly Cloudy	70° / 49°	61° / 42°	63° / 47°
Barcelona, Spain Partly Cloudy	73° / 57°	73° / 54°	73° / 56°

* Severe weather alert

City or ZIP Go

Yahoo! Noticias: Foto de Portada Options

PESHAWAR, Pakistán (AFP) - Ataque suicida y primera victoria en ofensiva en el sur de Pakistán

» Más

Yahoo! Mail Preview Options

fpjunqueira

Lufthansa	03:37 am
Travel information for your flight on 28/Oct to Frankfurt/Main, Mr Junqueira	
word@m-w.com	Oct 24
gruntle: M-W's Word of the Day	
Ryan Schmidt	Oct 23
Barra in the Times	
Travelocity Deals	Oct 23
Save up to 40%, plus affordable Asia vacations	
Economist.com	Oct 23
New on Economist.com - 23rd October 2009	

Compose Refresh

Yahoo! News: Most Viewed Options

Church janitor charged in slaying of NJ priest (AP) - 2 hours ago

AP - Authorities investigating the slaying of a priest arrested the church janitor Saturday, alleging he stabbed the cleric 32 times with a kitchen knife after arguing with him in the rectory.



Yahoo! Portal

The screenshot shows the Yahoo! Portal homepage with various sections and annotations. The top navigation bar includes links for Web, Images, Video, Local, Shopping, and more. A search bar with a 'Web Search' button is located at the top right. Below the navigation bar, there are tabs for 'My Front Page', 'The Best of My Yahoo! NEW', and 'New Tab'. A user profile 'Hi, Flavio' is visible with links for 'Sign Out', 'Tips', and 'Help'. A row of links includes 'Add Content', 'Change Appearance', 'More Options', and 'My Yahoo! Blog: It's Y!ou'. The main content area is divided into several sections: 'Personal Assistant' with links for Mail, Horoscope, Stocks, Weather, Lottery, and Sports; 'Message Center'; 'Weather' section showing '57°F Mostly Cloudy' for Philadelphia, PA, and a table for Today, Tomorrow, and Monday; 'Yahoo! Noticias: Foto de Portada' with a headline about a suicide attack in Pakistan; 'Yahoo! Mail Preview' showing a list of emails from 'fpjunqueira'; and 'Yahoo! News: Most Viewed' with a headline about a church janitor charged in the slaying of a priest. Annotations with purple arrows point from the following labels to specific elements on the page: 'Search' points to the search bar; 'E-mail' points to the Mail link in the Personal Assistant; 'Finance' points to the Stocks link in the Personal Assistant; 'Weather' points to the 57°F weather display; and 'News' points to the headline about the church janitor.

MY YAHOO!

Web Images Video Local Shopping more

Quicklinks My Front Page The Best of My Yahoo! NEW New Tab

Hi, Flavio Sign Out Tips Help

Add Content Change Appearance More Options My Yahoo! Blog: It's Y!ou

Personal Assistant Options

Mail Horoscope Stocks Weather Lottery Sports

Message Center Options

Weather Options

Compact Classic Full

57°F Mostly Cloudy

Location Today Tomorrow Monday

Philadelphia, PA* Mostly Cloudy 70° / 49° 61° / 42° 63° / 47°

Barcelona, Spain Partly Cloudy 73° / 57° 73° / 54° 73° / 56°

* Severe weather alert

City or ZIP Go

Yahoo! Noticias: Foto de Portada Options

PESHAWAR, Pakistán (AFP) - Ataque suicida y primera victoria en ofensiva en el sur de Pakistán

Más

Yahoo! Mail Preview Options

fpjunqueira

Lufthansa 03:37 am

Travel information for your flight on 28/Oct to Frankfurt/Main, Mr Junqueira

word@m-w.com Oct 24

gruntle: M-W's Word of the Day Oct 23

Ryan Schmidt Oct 23

Barra in the Times

Travelocity Deals Oct 23

Save up to 40%, plus affordable Asia vacations

Economist.com Oct 23

New on Economist.com - 23rd October 2009

Compose Refresh

Yahoo! News: Most Viewed Options

Church janitor charged in slaying of NJ priest (AP) - 2 hours ago

AP - Authorities investigating the slaying of a priest arrested the church janitor Saturday, alleging he stabbed the cleric 32 times with a kitchen knife after arguing with him in the rectory.

Search

E-mail

Finance

Weather

News



Yahoo!: Workload generated



Yahoo!: Workload generated

The screenshot shows the Yahoo! homepage interface. At the top, there's a navigation bar with links for Web, Images, Video, Local, Apps, and More. Below this is the Yahoo! logo and a search bar with a 'Web Search' button. A user is logged in as 'Hi, Flavio' with options to 'Sign Out' or 'Page Options'. The left sidebar lists 'YAHOO! SITES' including Mail, Autos, Dating, Finance, Flickr, Games, Health, Horoscopes, Jobs, Messenger, Movies, News, omg!, Real Estate, Shine, Shopping, Sports, Travel, TV, and Weather (72°F). The main content area features a 'TODAY - October 09, 2011' section with a featured article 'Things that claim to work, but don't' about walk signals. Below this are four small thumbnails: '5 things that don't work', 'Paris Jackson dresses like dad', 'Classy gesture after MLB win', and 'Glamour guys in big NFL game'. To the right is a 'TRENDING NOW' list with 10 items, including Nancy Shevell, Michelle Williams, Ben Stiller, Casey Anthony, Consumer credit, Evangeline Lilly, UFC 136, Sesame Street, H1N1 treatment, and Home ownership. Further right is a 'Yahoo! Search' section with a 'What's Happening Now' grid showing Anne Hathaway, House, and Breakfast Recipes. Below that is a 'VIDEO PICKS' section with a video titled 'Haunted house camera captures faces of fear'. At the bottom, there's a 'NEWS' section with headlines about missing K.C. baby, Paul McCartney, and former Weezer bassist Mikey Welsh. The footer includes 'MY FAVORITES' with a Facebook link and a 'More: News | Popular | Photos' link.



Yahoo!: Workload generated

The image is a screenshot of the Yahoo! homepage as it appeared on October 9, 2011. The page features the classic Yahoo! layout with a top navigation bar, a search bar, and various content sections. A large purple text box is overlaid on the left side of the page, containing the text: "Yahoo! Homepage: 4.4 billion page views a month".

Yahoo! Homepage: 4.4 billion page views a month

The screenshot shows the following elements:

- Top Navigation:** Web, Images, Video, Local, Apps, More.
- Search Bar:** A search bar with the text "Web Search" and a "Go" button.
- User Area:** "Hi, Flavio" with a "Sign Out" link and "Page Options".
- Left Sidebar:** "Y! Schweiz", "My Yahoo!", "Make Y! your homepage". Below this is a "YAHOO! SITES" section with links to Mail, News, omg!, Real Estate, Shine, Shopping, Sports, Travel, TV, and Weather (72°F).
- Main Content Area:**
 - TODAY - October 09, 2011:** A section with a large image and text: "Does the walk signal really change because you press the button, or is it a placebo? The truth behind it »".
 - 5 things that don't work:** A list of five items: "What is a placebo?", "Unhealthy health drinks", "Do toning shoes work?", "Paris Jackson dresses like dad", "Classy gesture after MLB win", "Glamour guys in big NFL game".
 - NEWS:** A section with headlines: "Parents of missing K.C. baby resume talking to police", "Paul McCartney arrives for wedding in central London", "Former Weezer bassist Mikey Welsh found dead in Chicago", "Hertz: Muslim workers can return if they follow break rules", "4 arrested for allegedly cutting hair, beard of Amish man", "Authorities say body found in missing Wash. man's SUV", "Wall Street Protest Spurs Online Conversation - N.Y. Times", "Wall St. protesters march Washington Sq. - New York Dai...", "Autistic LI boy steals truck - New York Post".
- Right Sidebar:**
 - TRENDING NOW:** A list of ten trending topics: 1. Nancy Shevell, 2. Williams, 3. Tony credit, 4. Evangeline Lilly, 5. UFC 136, 6. Sesame Street, 7. H1N1 treatment, 8. Home ownership.
 - AdChoices:** A section with a "Search Web" bar and a "Go" button.
 - VIDEO PICKS:** A section with a video player and a "Go to Video" button. The video is titled "Haunted house camera captures faces of fear".



Yahoo!: Workload generated

The image displays two versions of the Yahoo! homepage. The left version is a full-width screenshot showing the 'YAHOO!' logo, navigation links (Web, Images, Video, Local, Apps, More), a search bar, and a sidebar with various services like Mail, Health, Horoscopes, Jobs, Messenger, Movies, News, omg!, Real Estate, Shine, Shopping, Sports, Travel, TV, and Weather. A purple text box is overlaid on the sidebar with the text 'Yahoo! H billion po'. The right version is a zoomed-in view of the search bar area, showing the 'Web Search' button and a link to 'Make Yahoo! your homepage'.



Yahoo!: Workload generated

The image displays two screenshots of the Yahoo! homepage to illustrate the search workload. The left screenshot shows the full homepage layout with a sidebar of various services. A purple text box is overlaid on the left side of this screenshot, stating: "Yahoo! H... billion po...". The right screenshot is a zoomed-in view of the search bar area. A purple text box is overlaid on the right side of this screenshot, stating: "Yahoo! Search: 2.7 billion queries a month".

Left Screenshot (Full Page):

- Navigation: Web, Images, Video, Local, Apps, More
- Search Bar: Web Search
- User: Hi, Flavio | Sign Out | Page Options
- Sidebar (Left): Y! Schweiz, My Yahoo!, Make Y! your homepage; YAHOO! SITES (Edit); Mail, Health, Horoscopes, Jobs, Messenger, Movies, News, omg!, Real Estate, Shine, Shopping, Sports, Travel, TV, Weather (72°F); MY FAVORITES (Edit); Facebook
- Main Content: TODAY, Does the because placebo, 5 things don't w, 1 - 4 of 24, NEWS (Parent, Paul M, Former, Hertz, 4 arres, Author, Wall St, Wall St, Autistic, MLB | updated)

Right Screenshot (Zoomed Search Bar):

- Navigation: Web, Images
- Search Bar: Search
- Link: [Y! Make Yahoo! your homepage](#)
- Footer: © 2011 Yahoo! | Page Tour | Privacy / Legal | About Our Ads | Submit Your Site



Yahoo!: Workload generated

The screenshot shows the Yahoo! News homepage. At the top, there's a navigation bar with links for Web, Images, Video, Local, Apps, and More. Below this is the Yahoo! logo and a search bar. The main header includes the 'YAHOO! NEWS' logo and a search bar. A secondary navigation bar lists categories: HOME, U.S., WORLD, BUSINESS, ENTERTAINMENT, SPORTS, TECH, POLITICS, SCIENCE, HEALTH, BLOGS, LOCAL, and POPULAR. Below this is another row of links: VIDEO, PHOTOS, GMA, LOCAL, ODD NEWS, COMICS, TRAVEL, OPINION, TRENDING NOW, VITALITY, WANDERLUST, WHO KNEW?, and WEATHER.

The main content area features a large article titled 'Unemployed seek protection against job bias' with a photo of a woman. To the right, there's a 'MOST POPULAR' section with a list of headlines: 'California allows college aid to illegal immigrants', 'Romney's Mormonism in focus at political meeting', 'Eric Cantor says Wall Street protesters are 'mobs' as Democrats offer support', 'Apple tribute logo a Web hit', 'Protesters want world to know they're just like us', 'Unemployed seek protection against job bias', and 'Ron Paul Wins Straw Poll; Cain 2nd'.

Below the main article, there's a 'FEATURED BLOGS' section with two entries: 'Chris Moody - The Ticket' with the headline 'Pastor who endorsed Perry on Friday says Romney is a cultist' and 'Zachary Roth - The Lookout' with the headline 'Spent, an online game, forces players to'.

At the bottom left, there's a 'Top Stories' section with a link to 'Syria warns countries not to recognize opposition'.

n: 2.7
s a month



Yahoo!: Workload generated

Web Images Video Local Apps More

YAHOO! NEWS

Hi, Flavio | Sign Out | Help

Make Y! home, help a school

Web Search

Search

Search Web

HOME U.S. WORLD BUSINESS ENTERTAINMENT SPORTS TECH POLITICS SCIENCE HEALTH BLOGS LOCAL POPULAR

VIDEO PHOTOS GMA LOCAL ODD NEWS COMICS TRAVEL OPINION TRENDING NOW VITALITY WANDERLUST WHO KNEW? WEATHER

NEW Your news. Now with friends.

Discover News based on what your friends are reading and publish your own reading activity. You have full control over what you publish.

To get started, first [Login with Facebook](#)

Unemployed seek protection against job bias

Anti-Gadhafi fighters make gains in Sirte

FEATURED BLOGS

Chris Moody - The Ticket

Pastor who endorsed Perry on Friday says Romney is a cultist Fri, Oct 7, 2011

Zachary Roth - The Lookout

Spent, an online game, forces players to

Syria warns countries not to recognize opposition AP - 2 mins 44 secs ago

Syria's foreign minister warned the international community Sunday not to recognize a new umbrella council formed by the opposition, threatening "tough measures" against any country that does so. [More »](#)

Yahoo! News: 88 million users in the US and 256 million users globally

n: 2.7
s a month



Yahoo! Infrastructure

- Lots of servers
- Lots of processes
- High volumes of data
- Highly complex systems
- ... and developers are mere mortals



Yahoo! Lockport Data Center



by amusingplanet via Flickr



Copyright by Peter E. Lee via Flickr



Copyright by Shamus O'Reilly via Flickr



Copyright by Shamus O'Reilly via Flickr

... and in computer systems?

- Locks
- Queues
- Leader election
- Group membership
- Barriers
- Configuration



... and in computer systems?

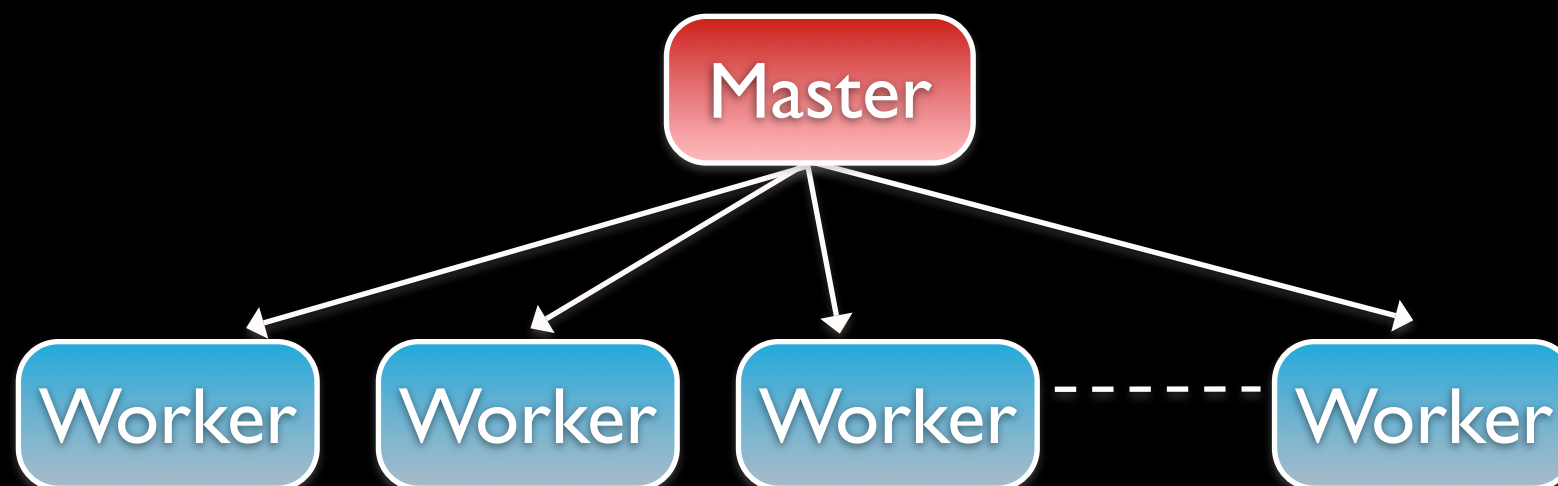
- Locks
- Queues
- Leader election
- Group membership
- Barriers
- Configuration

Require knowledge
of complex protocols



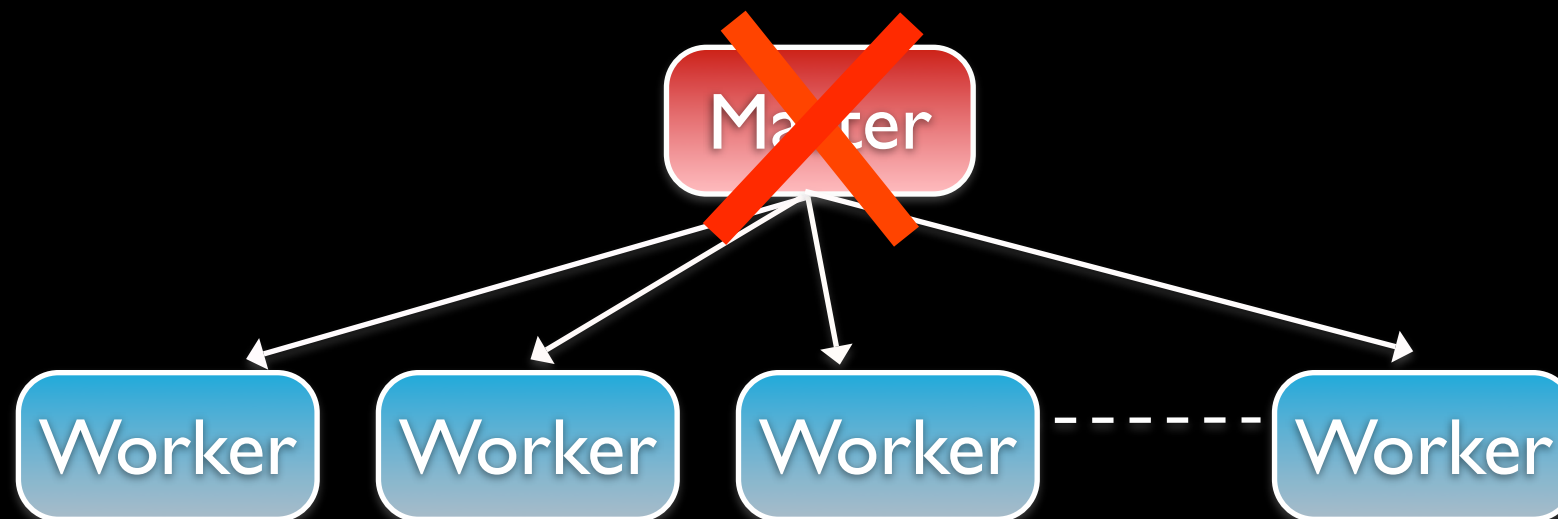
A simple model

- Work assignment
 - ✓ Master assigns work
 - ✓ Worker executes task assigned by master



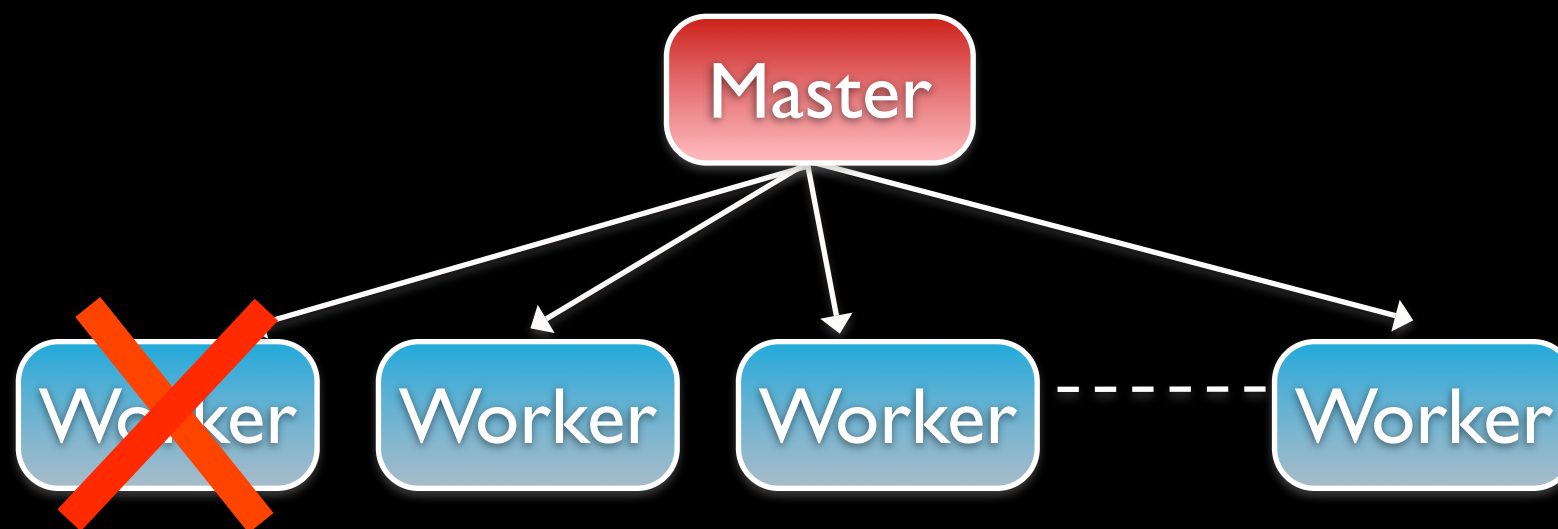
Master crashes

- ✓ Single point of failure
- ✓ No work assigned
- ✓ Need to select a new master



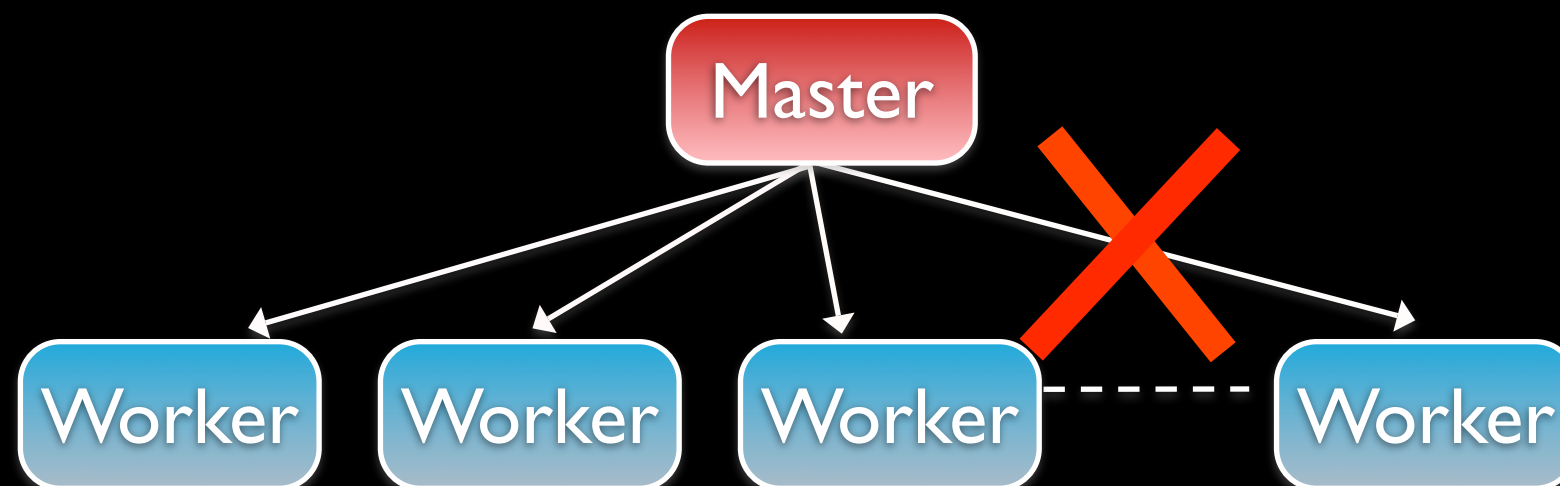
Worker crashes

- ✓ Not as bad... Overall system still works
- ✓ Some tasks won't be executed
- ✓ Need the ability to reassign tasks



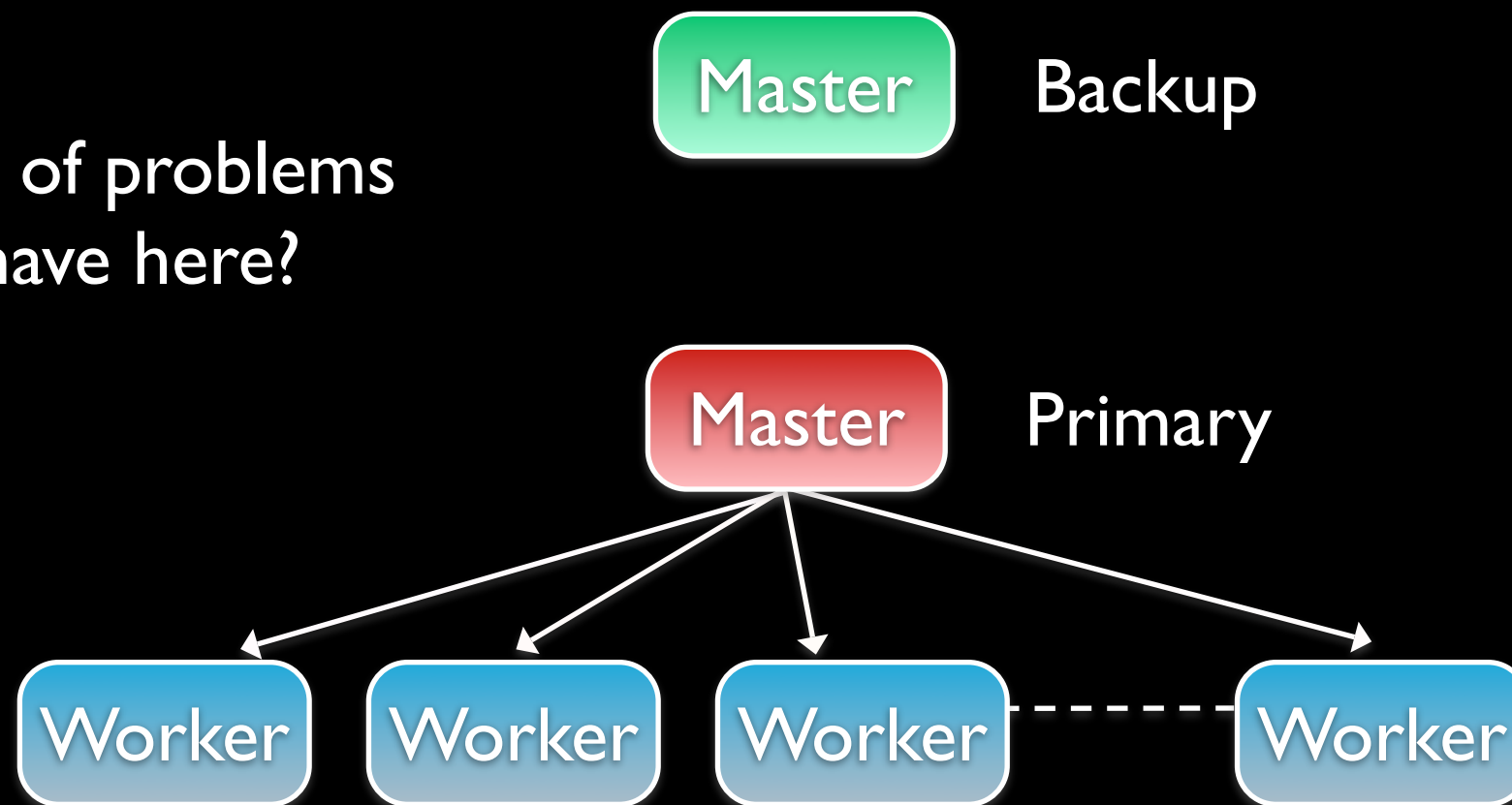
Worker does receive assignment

- ✓ Same problem, tasks don't get executed
- ✓ Need to guarantee that worker receives assignment

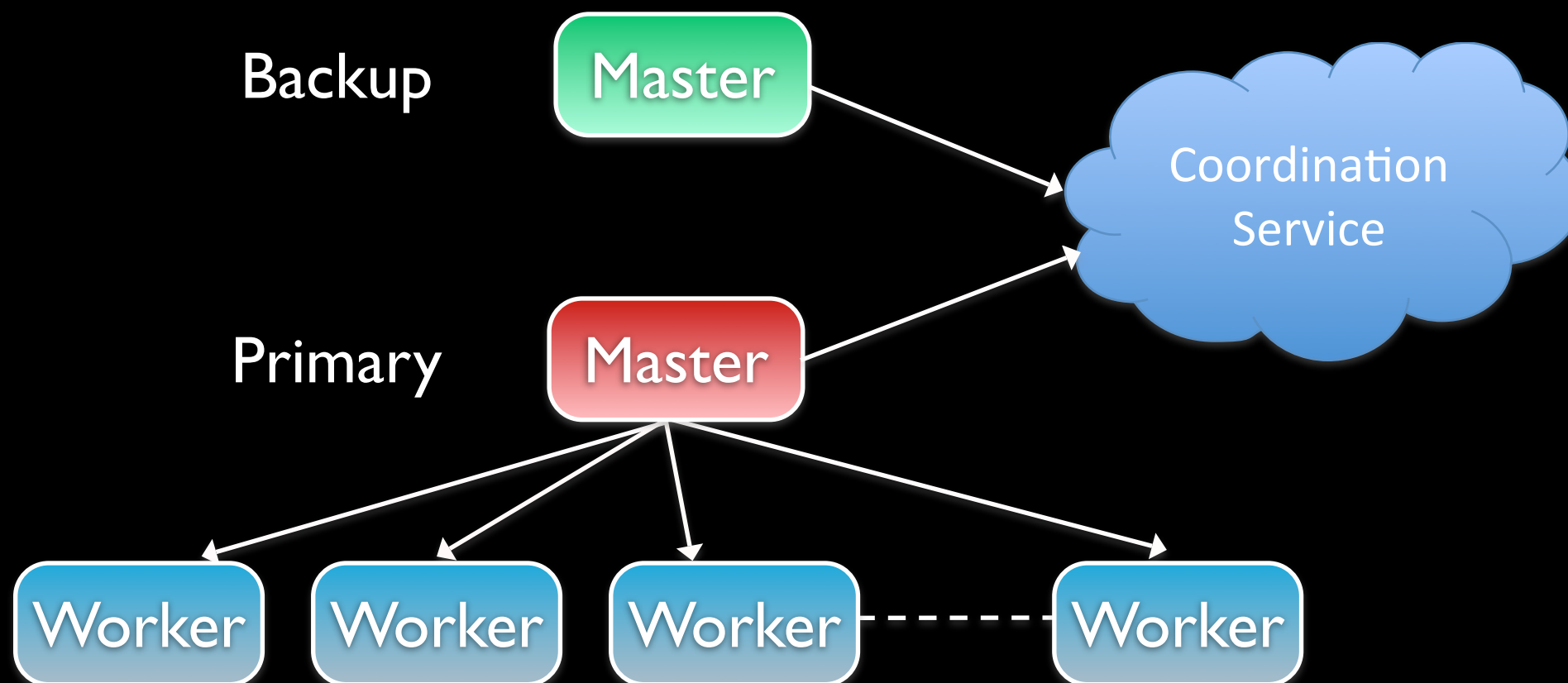


Fault-tolerant distributed system

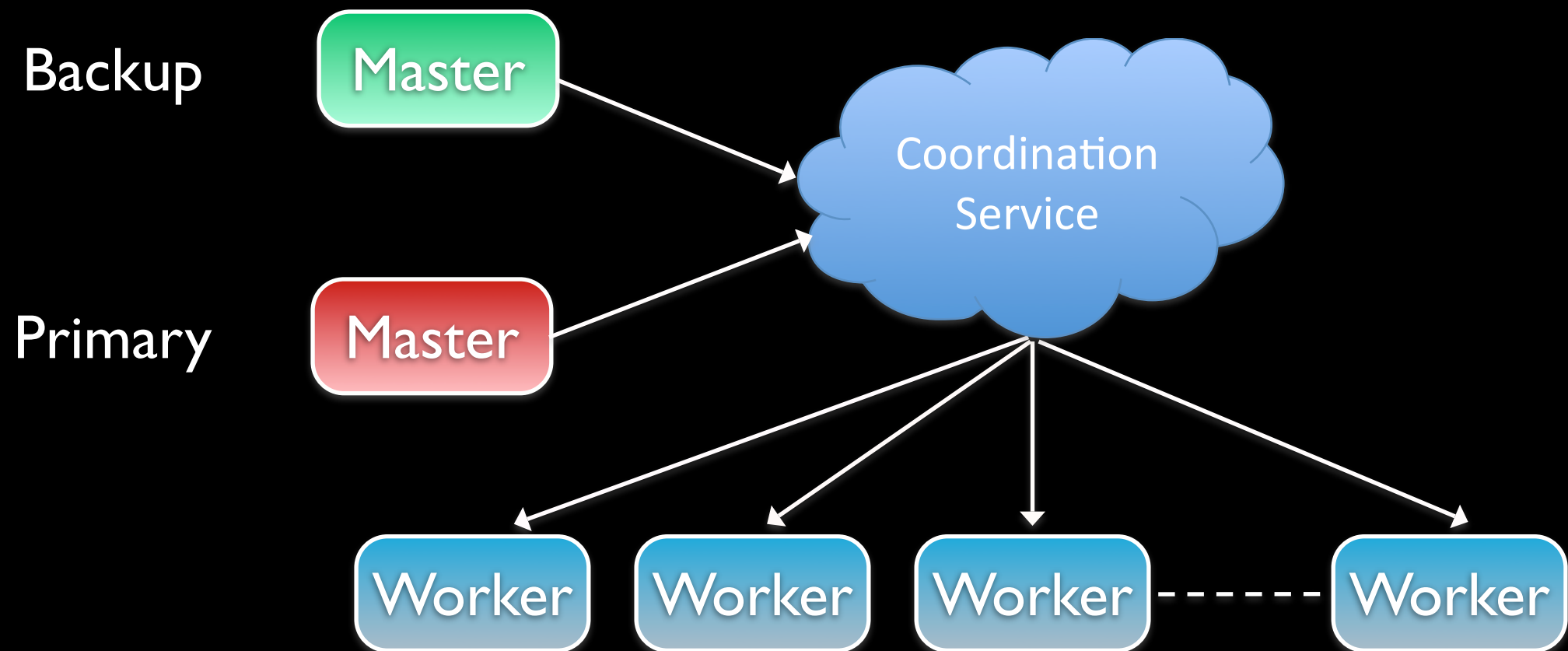
What kinds of problems
can we have here?



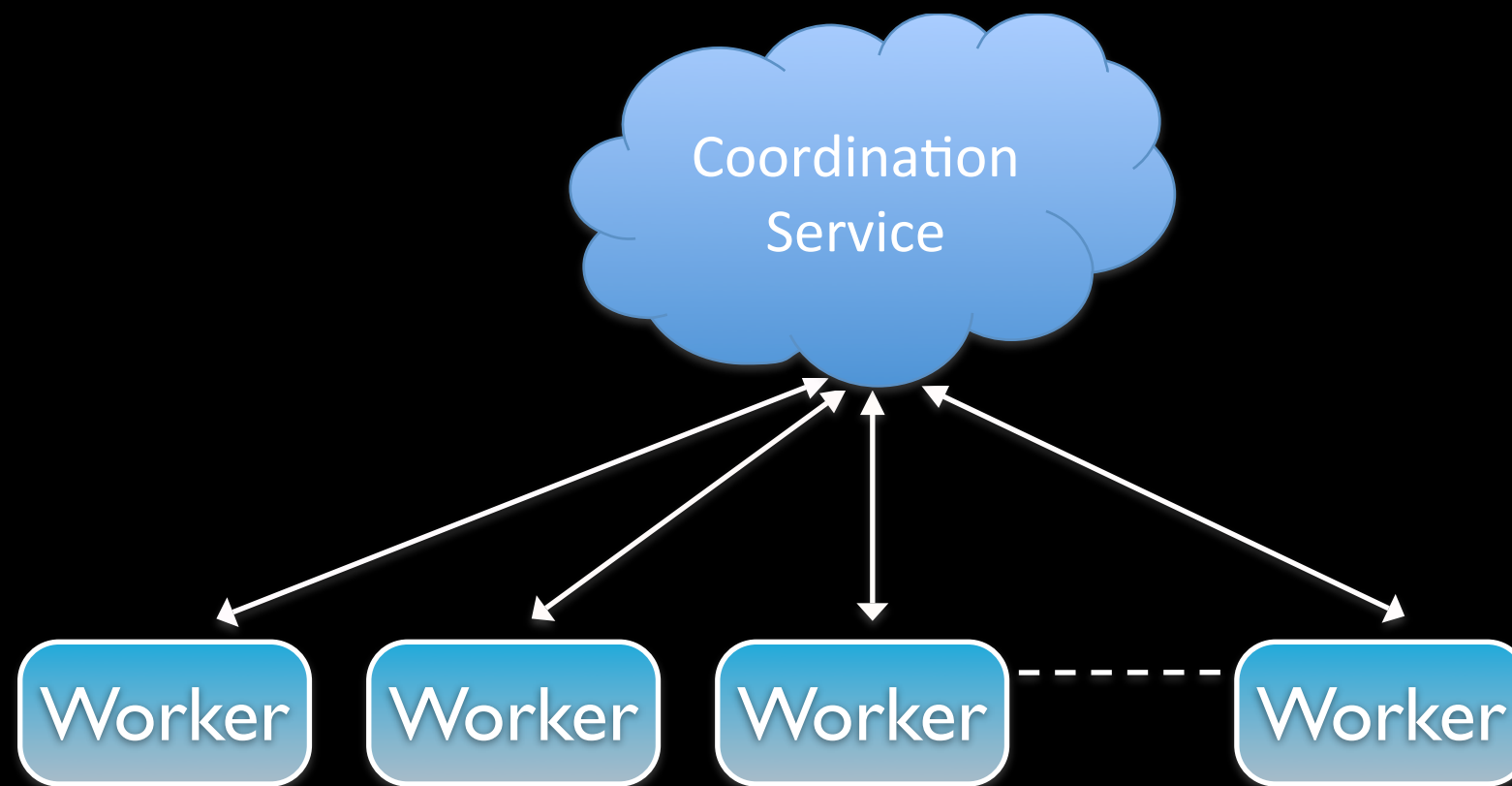
Fault-tolerant distributed system



Fault-tolerant distributed system



Fully distributed



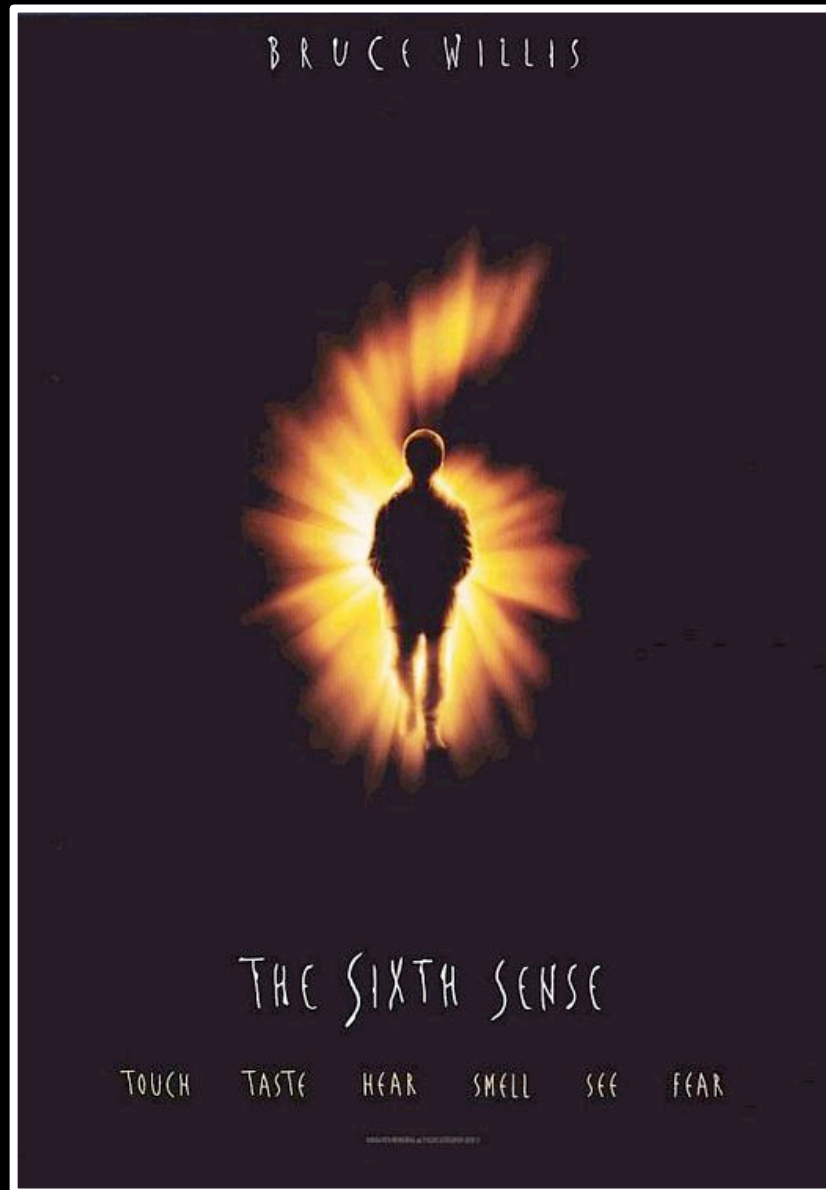
Fallacies of distributed computing

1. The network is reliable
2. Latency is zero
3. Bandwidth is infinite
4. Network is secure
5. Topology doesn't change
6. There is one administrator
7. Transport cost is zero
8. Network is homogeneous

Peter Deutsch, <http://blogs.sun.com/jag/resource/Fallacies.html>



One more fallacy



9. You know who is alive



Why is it difficult?

- FLP impossibility result

- ✓ Asynchronous systems
- ✓ Impossible if a single process can crash

Fischer, Lynch, Paterson, ACM PODS, 1983

- According to Herlihy we do need consensus

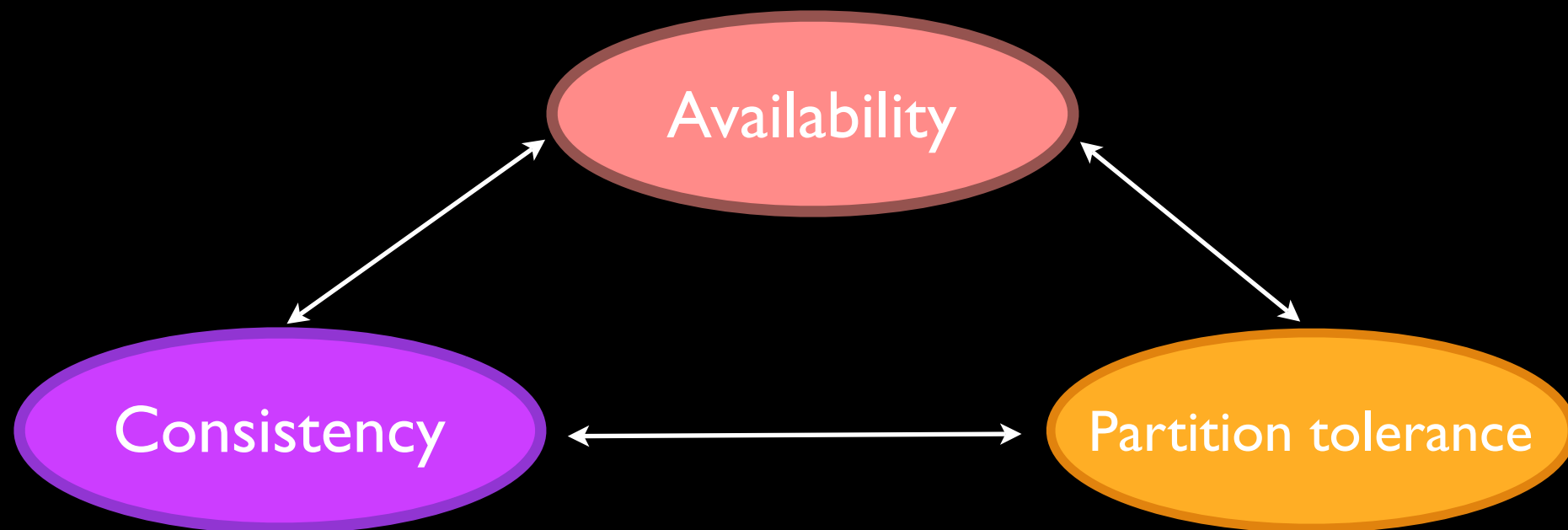
- ✓ Wait-free synchronization
- ✓ Wait-free: Operations complete in a finite number of steps
- ✓ Equivalent to solving consensus for n processes

Herlihy, ACM TOPLAS, 1991



Why is it difficult?

- CAP Principle
 - ✓ Can't have availability, consistency, and partition tolerance



The case for a coordination service

- Many fallacies to stumble upon
- Many impossibility results
- Several common requirements across applications
 - ✓ Duplicating is bad
 - ✓ Duplicating poorly is even worse
- Coordination service
 - ✓ Implement it once and well



Current systems

- Google Chubby

- ✓ Lock service

Burrows, USENIX OSDI, 2006

- Microsoft Centrifuge

- ✓ Lease service

Adya et al., USENIX NSDI, 2010

- Apache ZooKeeper

- ✓ Coordination kernel

- ✓ Initially contributed by Yahoo!

Hunt et al., USENIX ATC, 2010



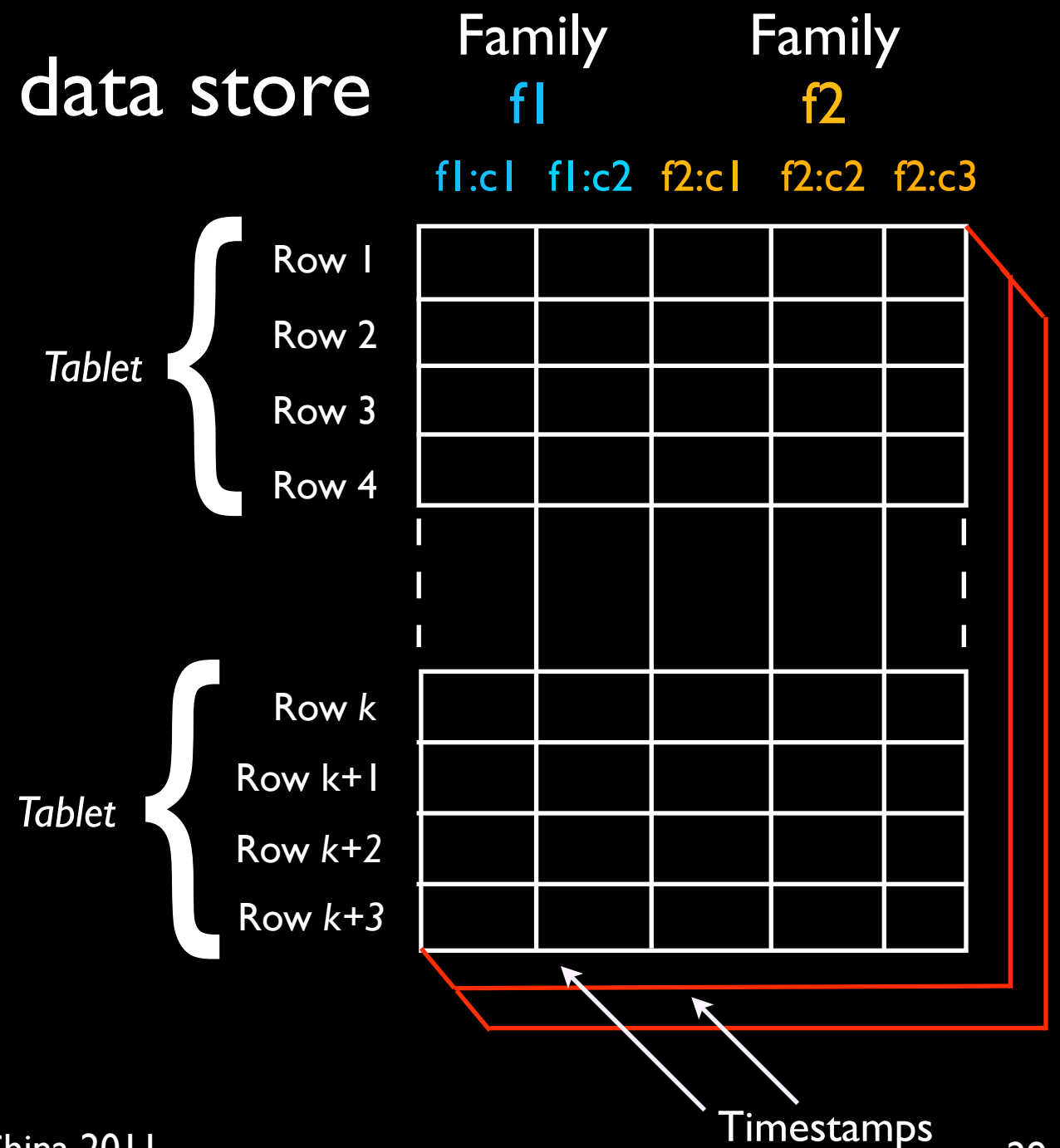
Example - Bigtable, Hbase

- Sparse column-oriented data store

- ✓ Tablet: Range of rows
- ✓ Unit of distribution

- Architecture

- ✓ Master
- ✓ Tablet servers



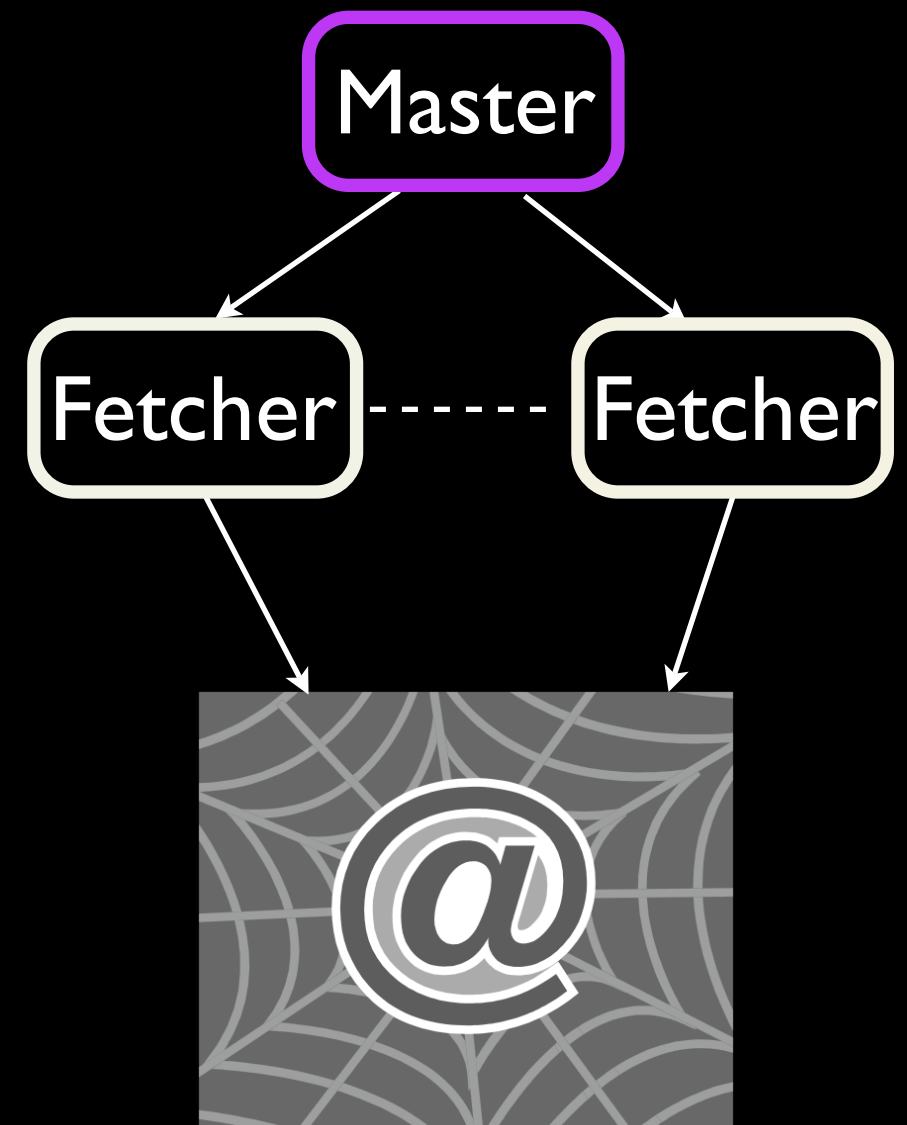
Example - Bigtable, Hbase

- Master election
 - ✓ Master crashes
- Metadata management
 - ✓ ACLs, Tablet metadata
- Rendezvous
 - ✓ Find tablet server
- Live tablet servers



Example - Web crawling

- Fetching service
 - ✓ Fetch web pages for search engines
- Master election
 - ✓ Assign work
- Metadata management
 - ✓ Politeness constraints
 - ✓ Shards
- Live workers



Example - Kafka, Pub/Sub

- Based on topics
- Topics are partitioned across brokers
- Consumer groups
 - ✓ Multiple consumers for a topic
- Coordination requirements
 - ✓ Metadata
 - ➡ Each message is consumed by a single consumer
 - ➡ Each partition is owned by a consumer
 - ✓ Crash detection



And more examples...

- Google File System
 - ✓ Master election
 - ✓ File system metadata
- Hedwig - Pub/Sub
 - ✓ Topic metadata
 - ✓ Topic assignment
 - ✓ Elasticity



At Yahoo!...

- Has been used for:
 - ✓ Fetching service
 - ✓ Manage workflows in Hadoop (e.g., feed ingestion)
 - ✓ Content optimization (distributed services)
 - ✓ ...
- Largest cluster I'm aware of
 - ✓ Around 5,000 - 10,000 clients



[illegible]

ZooKeeper introduction

- Coordination kernel
 - ✓ Does not export concrete primitives
 - ✓ Recipes to implement primitives
- File-system based API
 - ✓ Manipulates small data nodes: *znodes*
 - ✓ State is a hierarchy of *znodes*

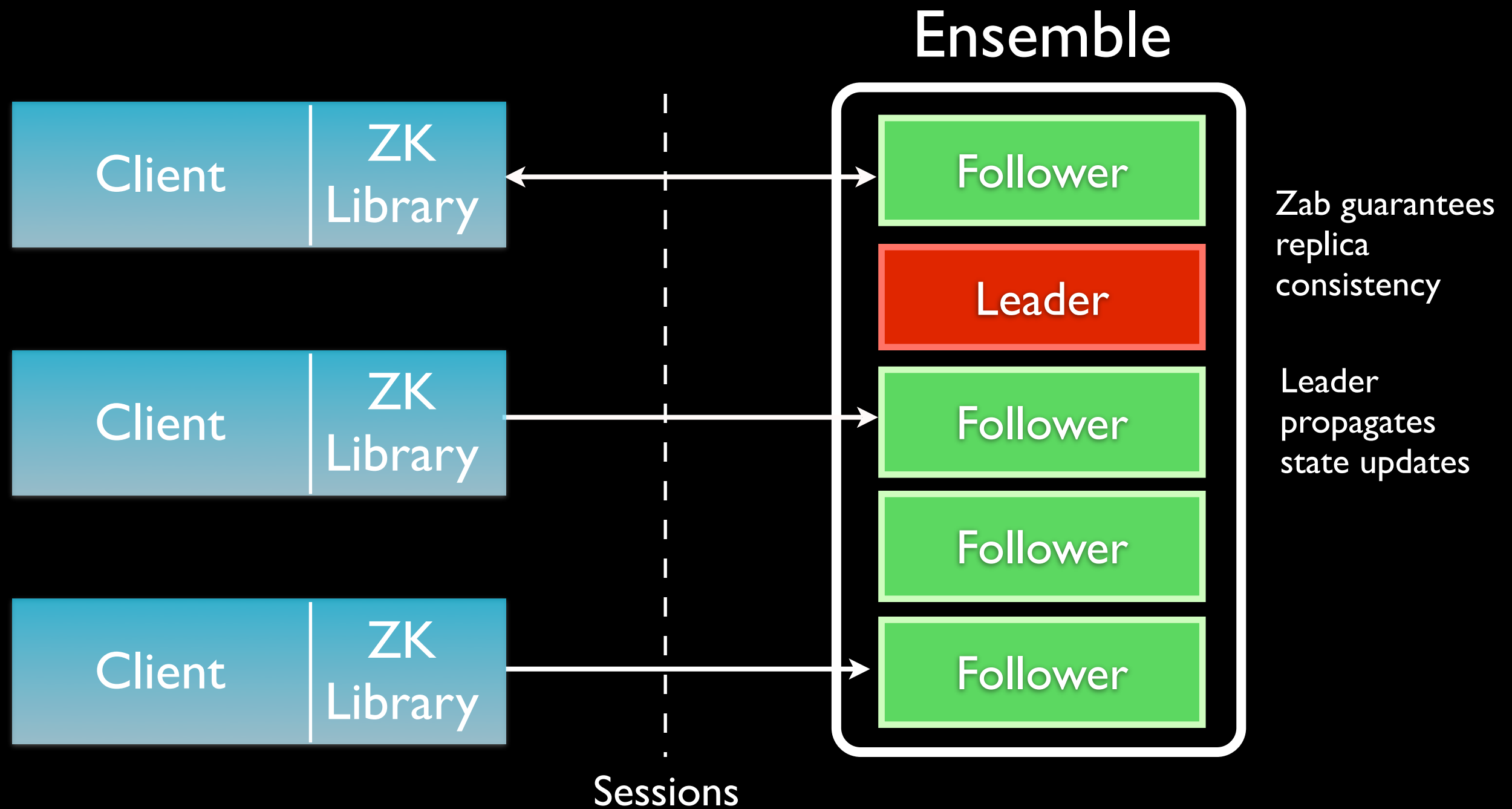


More introduction

- Stores database in memory
 - ✓ Handles high load
 - ✓ Handy when communicating with a large number of processes
- Single data-center applications, originally
 - ✓ Some cases of cross-colo deployments



ZooKeeper: Design



What's difficult here?

- Electing a leader
 - ✓ All live processes are potential leaders
 - ✓ Communication pattern is arbitrary
- Replicating the state
 - ✓ Zab: a high-performance broadcast protocol
 - ✓ Enables multiple outstanding operations



What do clients see?

- Semantics of sessions
- Prefix of operations are executed
- Upon a disconnection
 - ✓ A server tries to contact another server
 - ✓ Before session expires: connect to new server
 - ✓ Server must have seen a higher transaction id

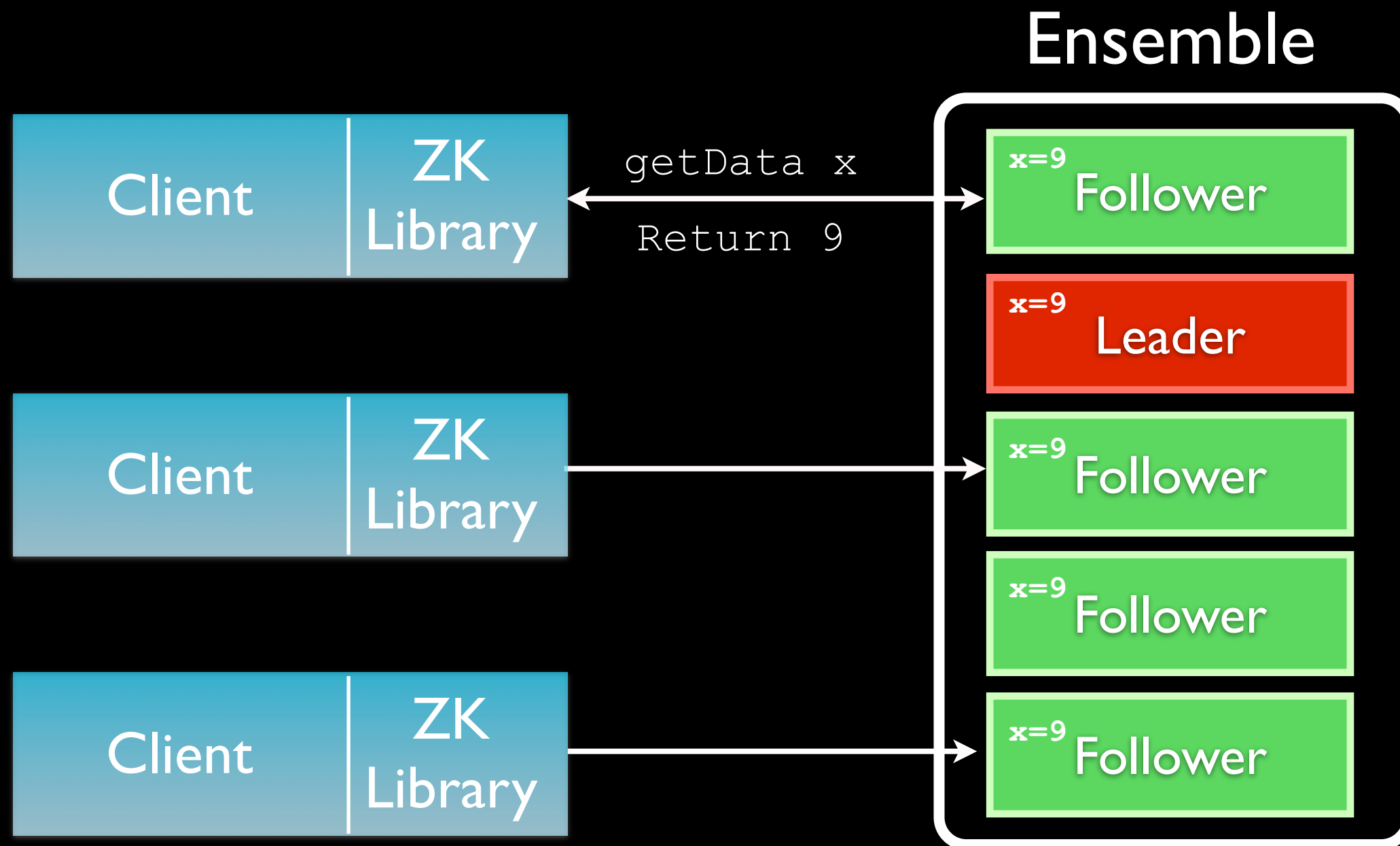


ZooKeeper API

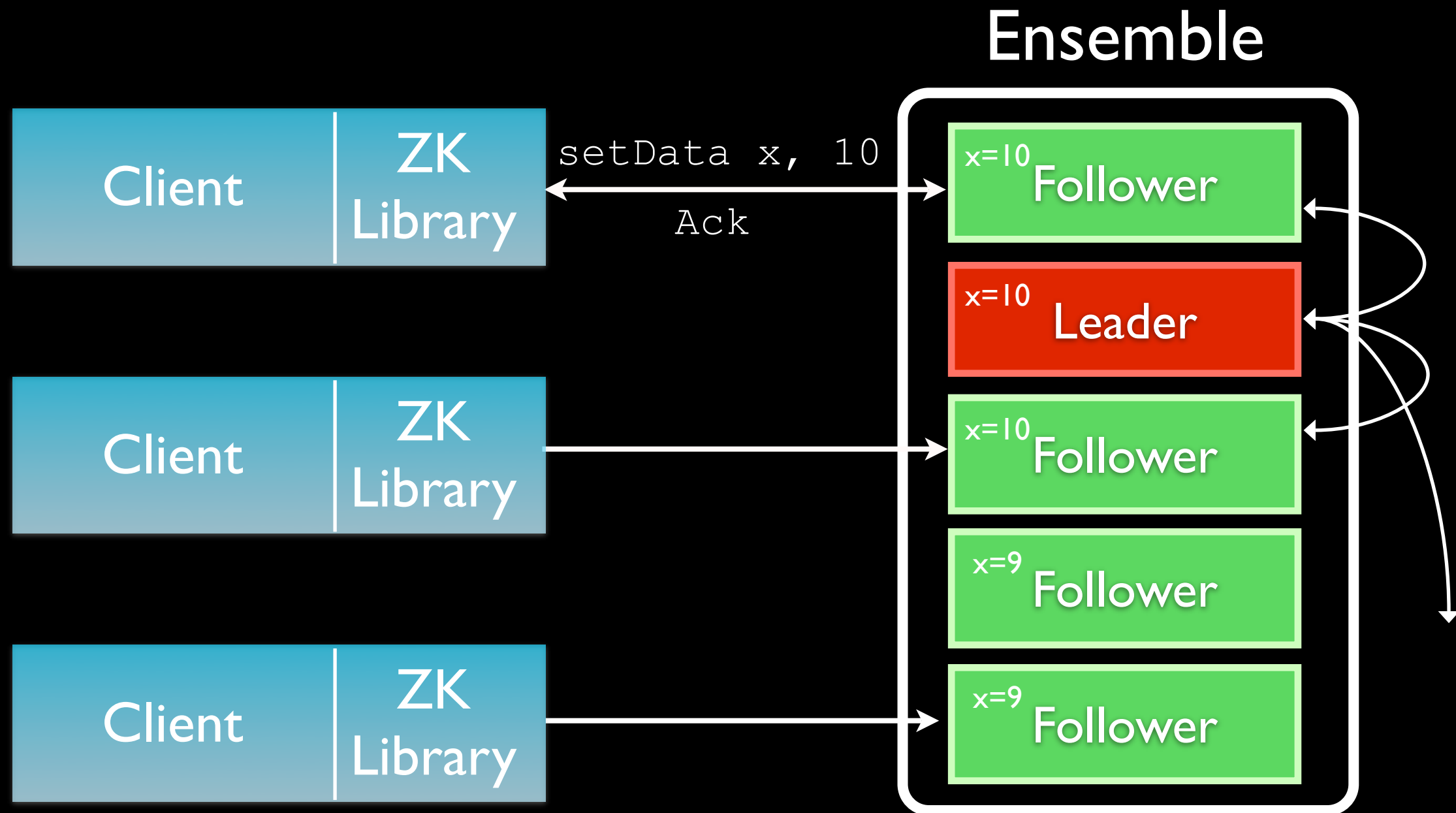
- Create znodes: `create`
 - ✓ Persistent, ephemeral, sequential
- Read and modify data: `getData`, `setData`
- Read the children of znode: `getChildren`
- Check if znode exists: `exists`
- Delete znode: `delete`



ZooKeeper: Reads

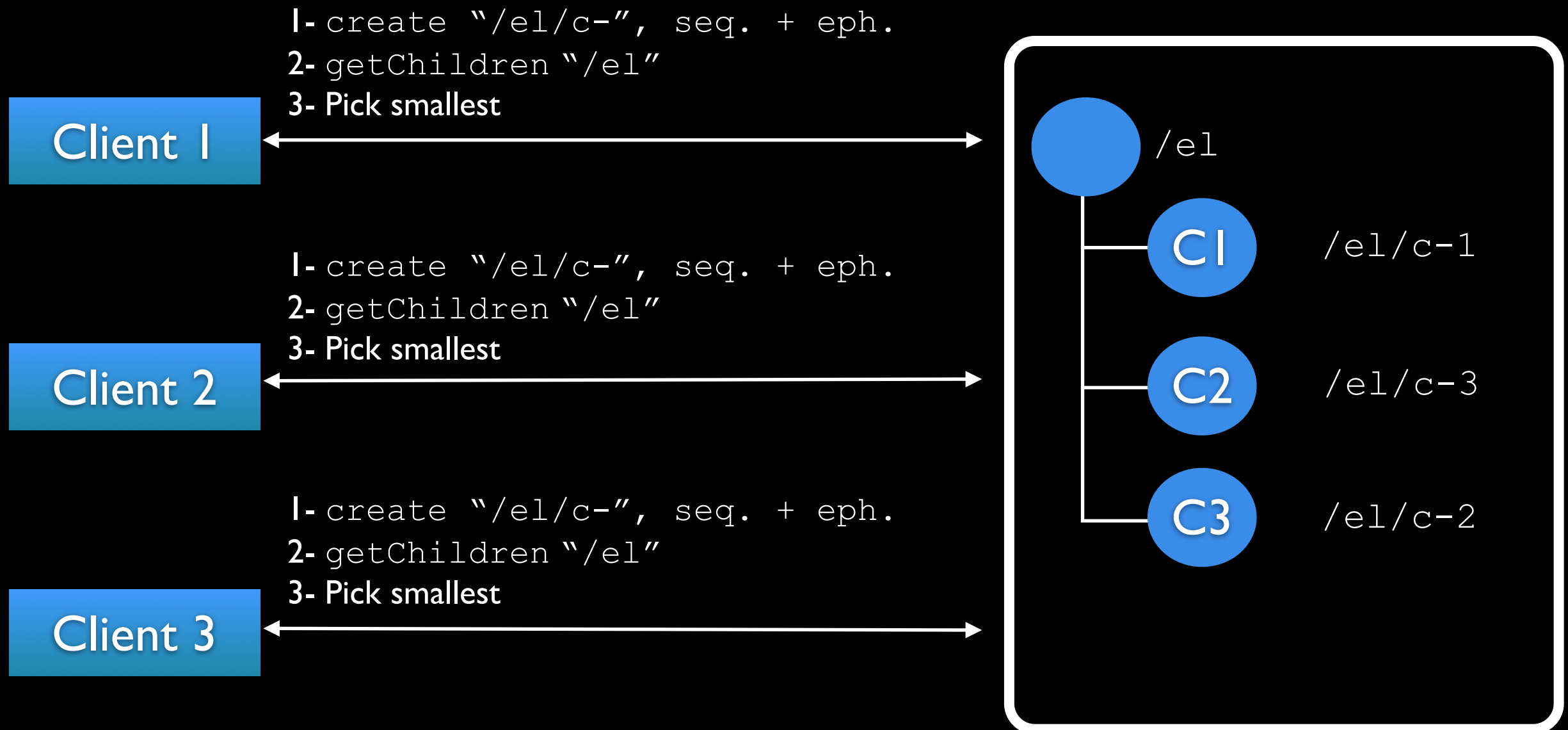


ZooKeeper:Writes



Example

Ensemble

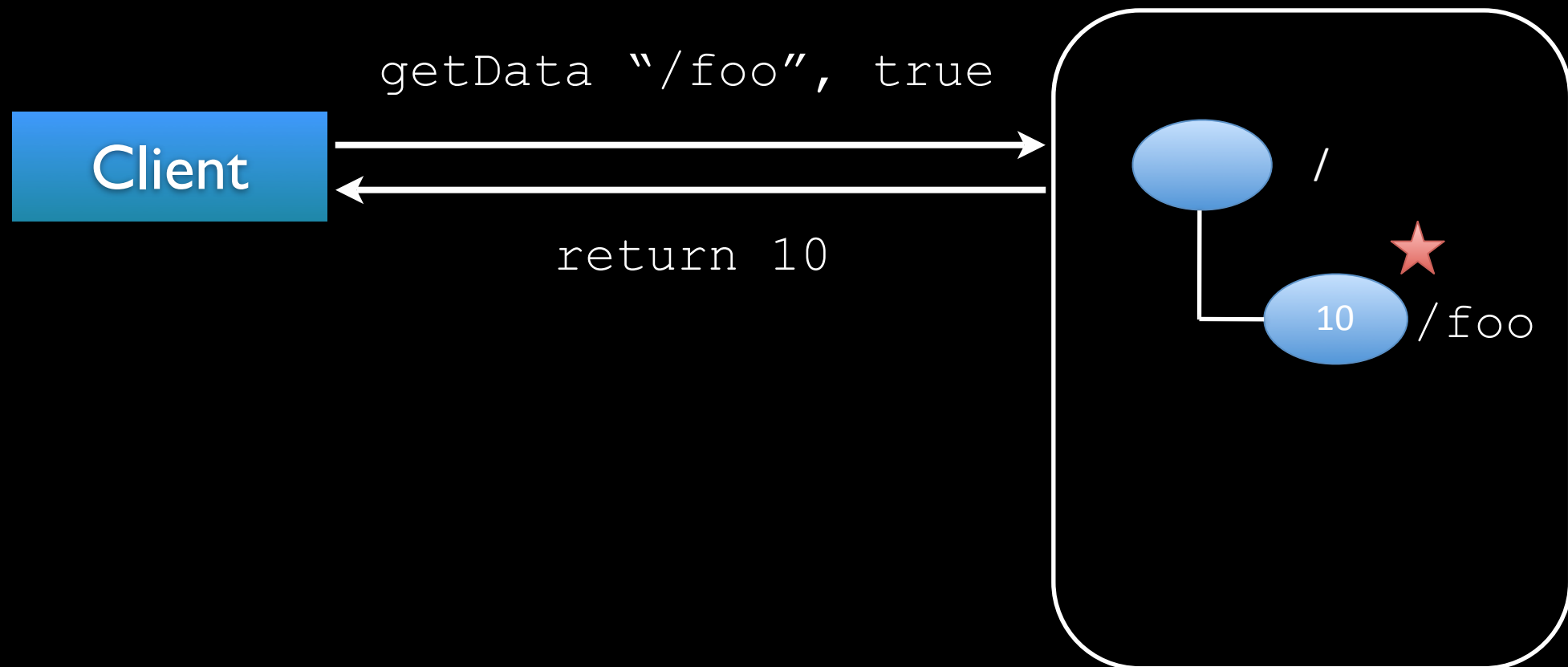


Znode changes

- Znode changes
 - ✓ Data is set
 - ✓ Node is created or deleted
 - ✓ *Etc...*
- Learn of znode changes
 - ✓ Set a *watch*
 - ✓ Upon change, client receives a *notification* before new updates



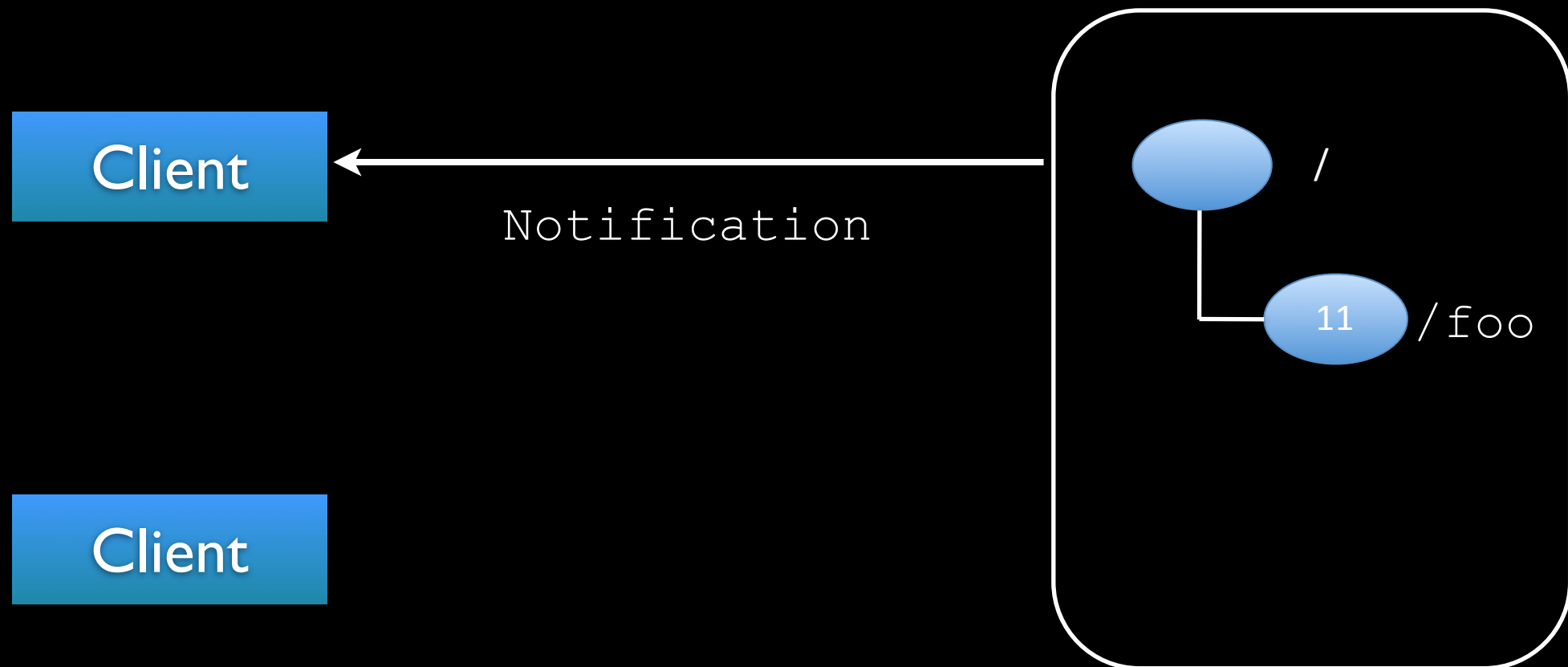
Watches



Watches



Watches



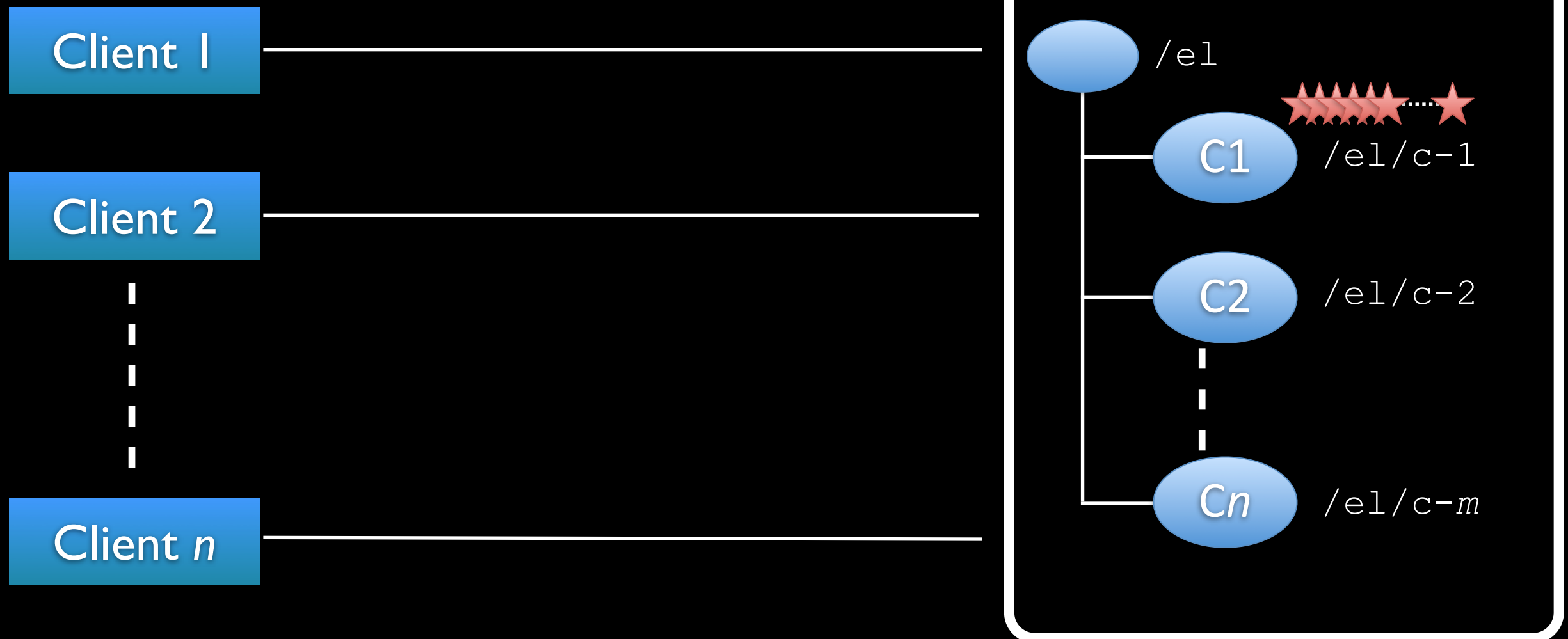
Watches, Locks, and the herd effect

- Herd effect
 - ✓ Large number of clients wake up simultaneously
- Undesirable effect
 - ✓ Load spikes

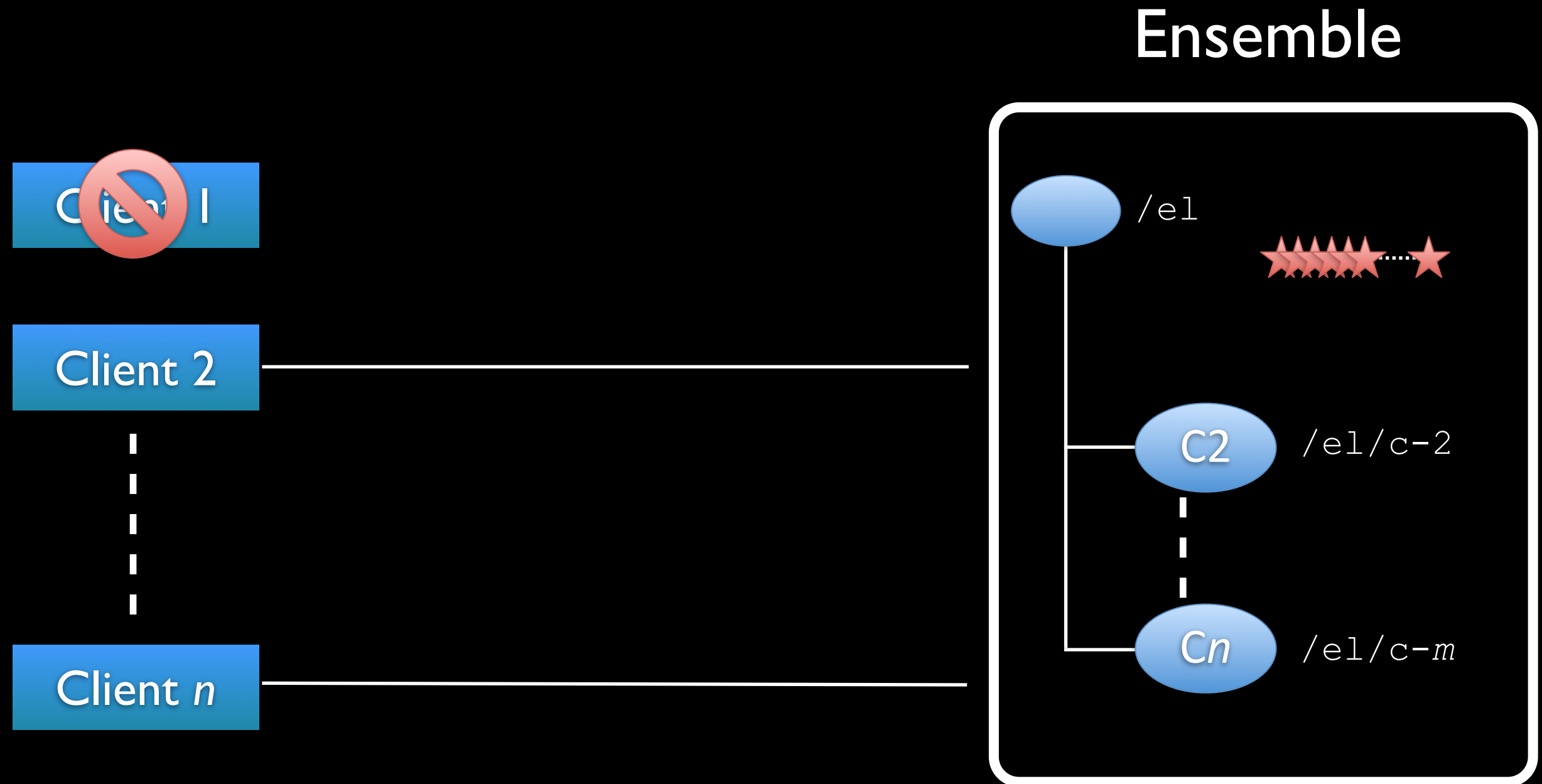


Watches, Locks, and the herd effect

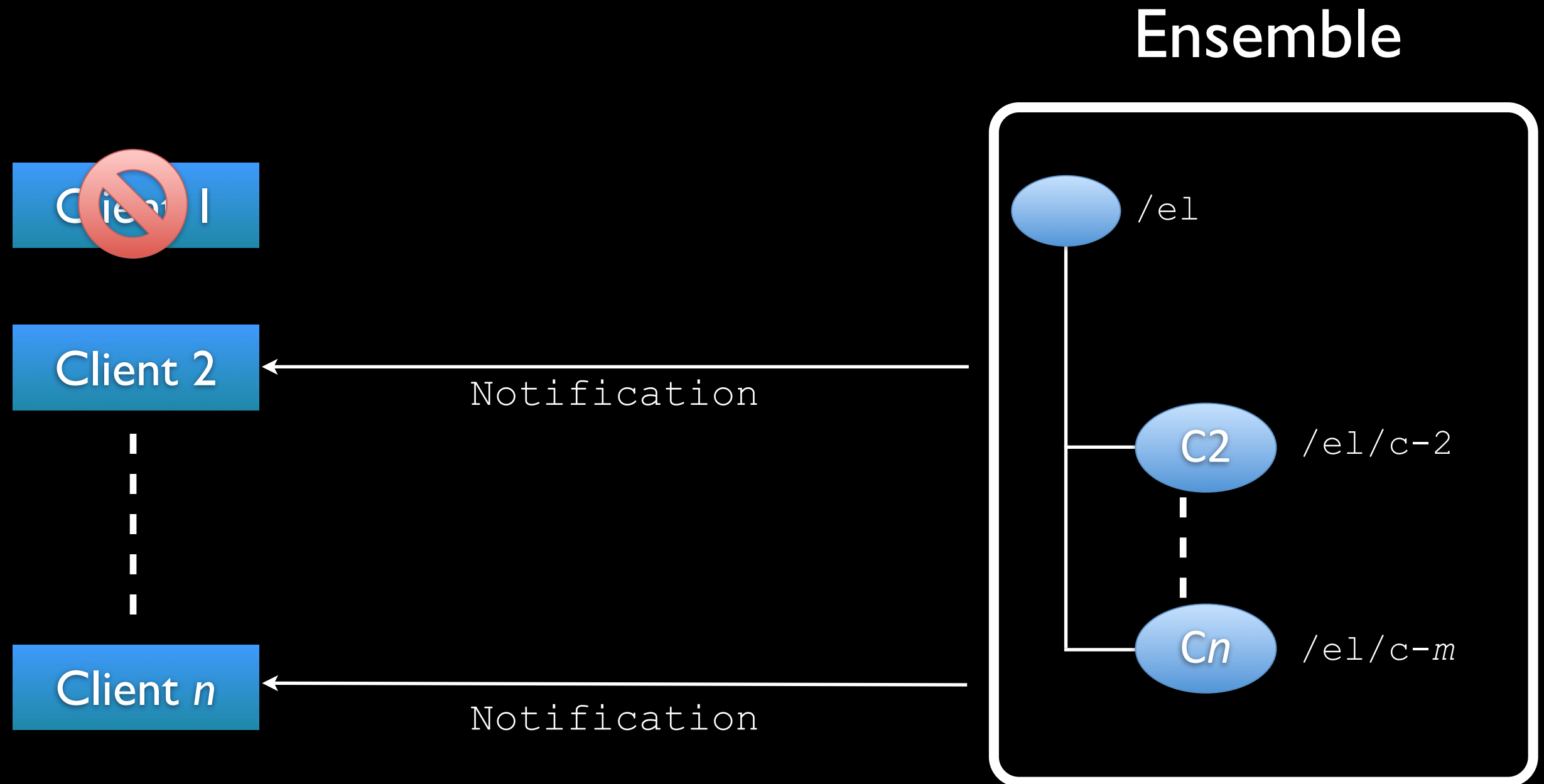
Ensemble



Watches, Locks, and the herd effect



Watches, Locks, and the herd effect



Watches, Locks, and the herd effect

- A solution: Use order of clients
 - ✓ Each client
 - ➡ Pick znode z preceding its own znode in the sequential order
 - ➡ Watch z
 - ✓ A single notification is generated upon a crash
- Works for locks
- Maybe not for leader election
 - ✓ One client is notified of a leader change





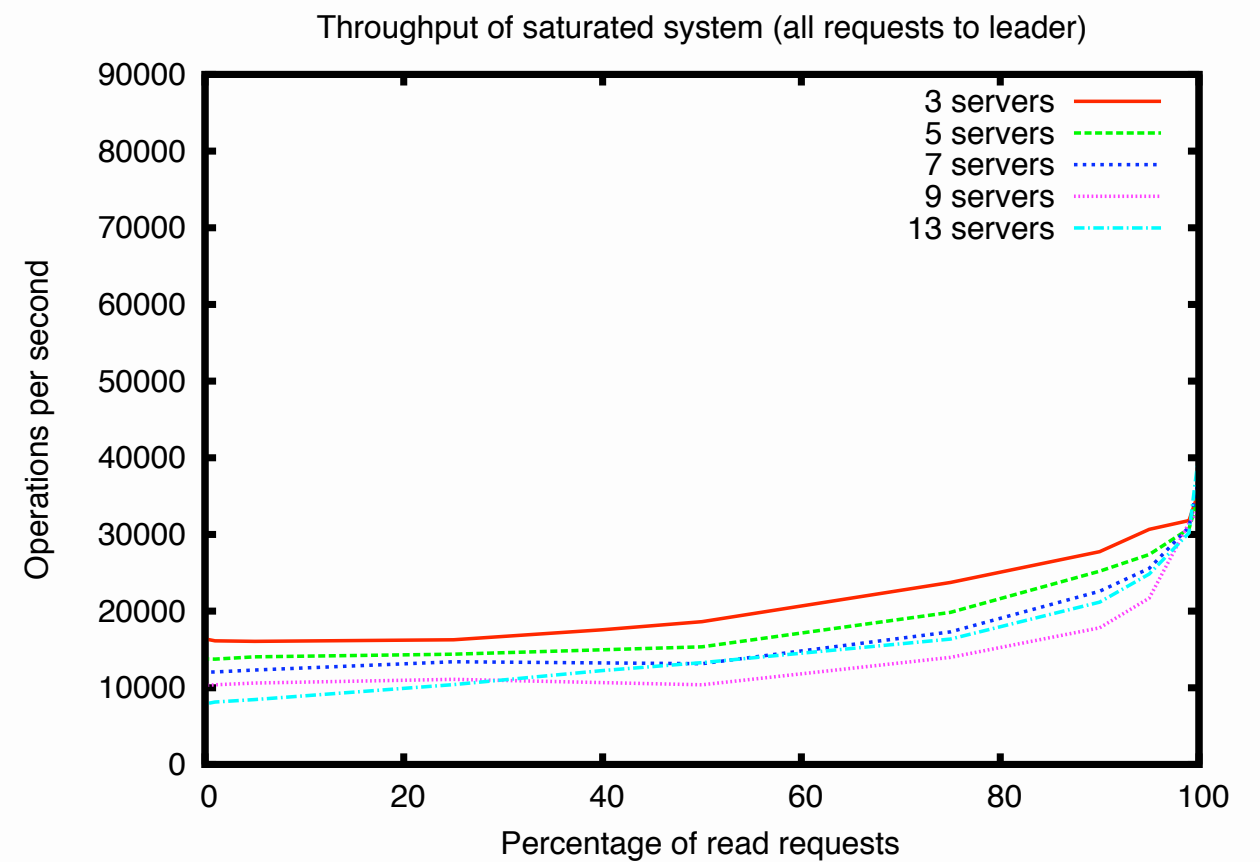
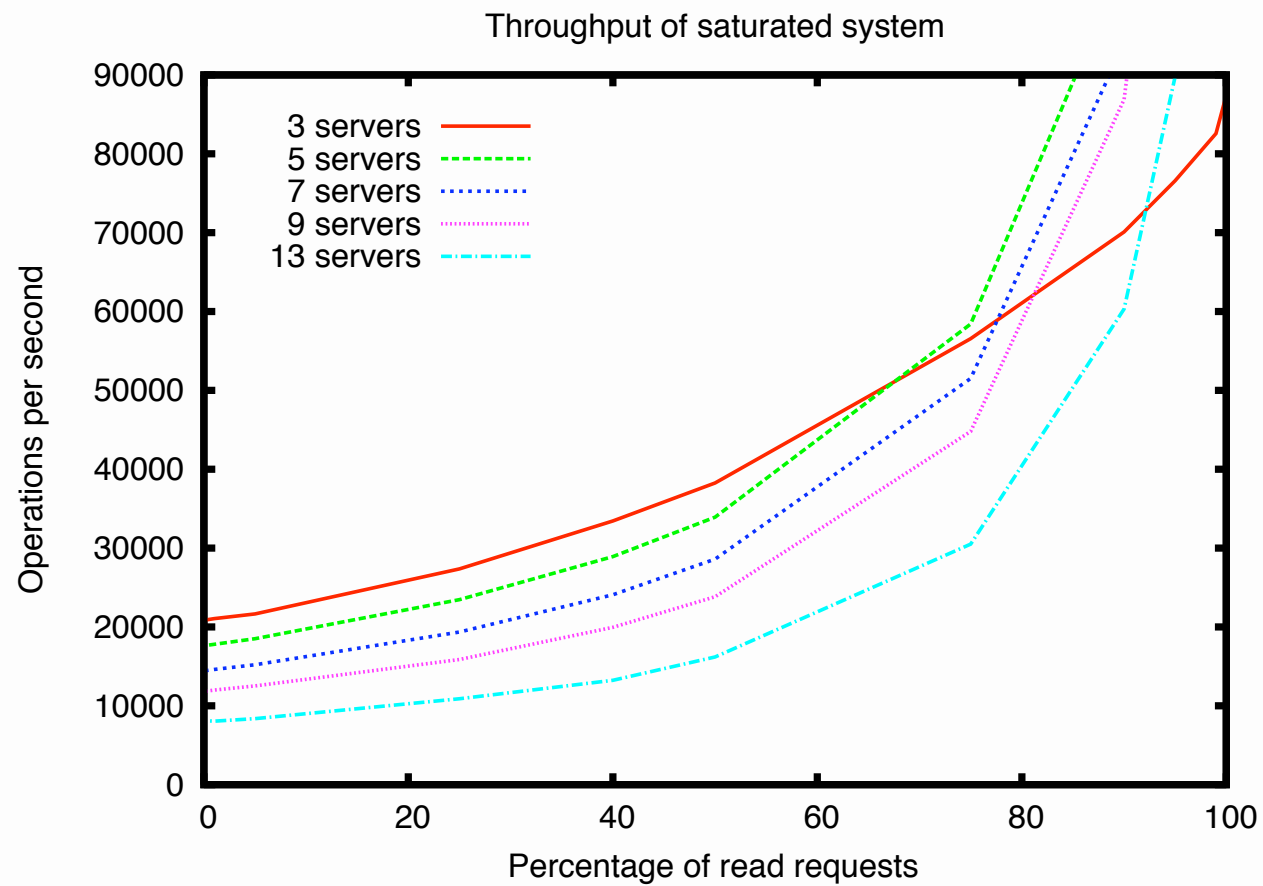
Evaluation

Evaluation

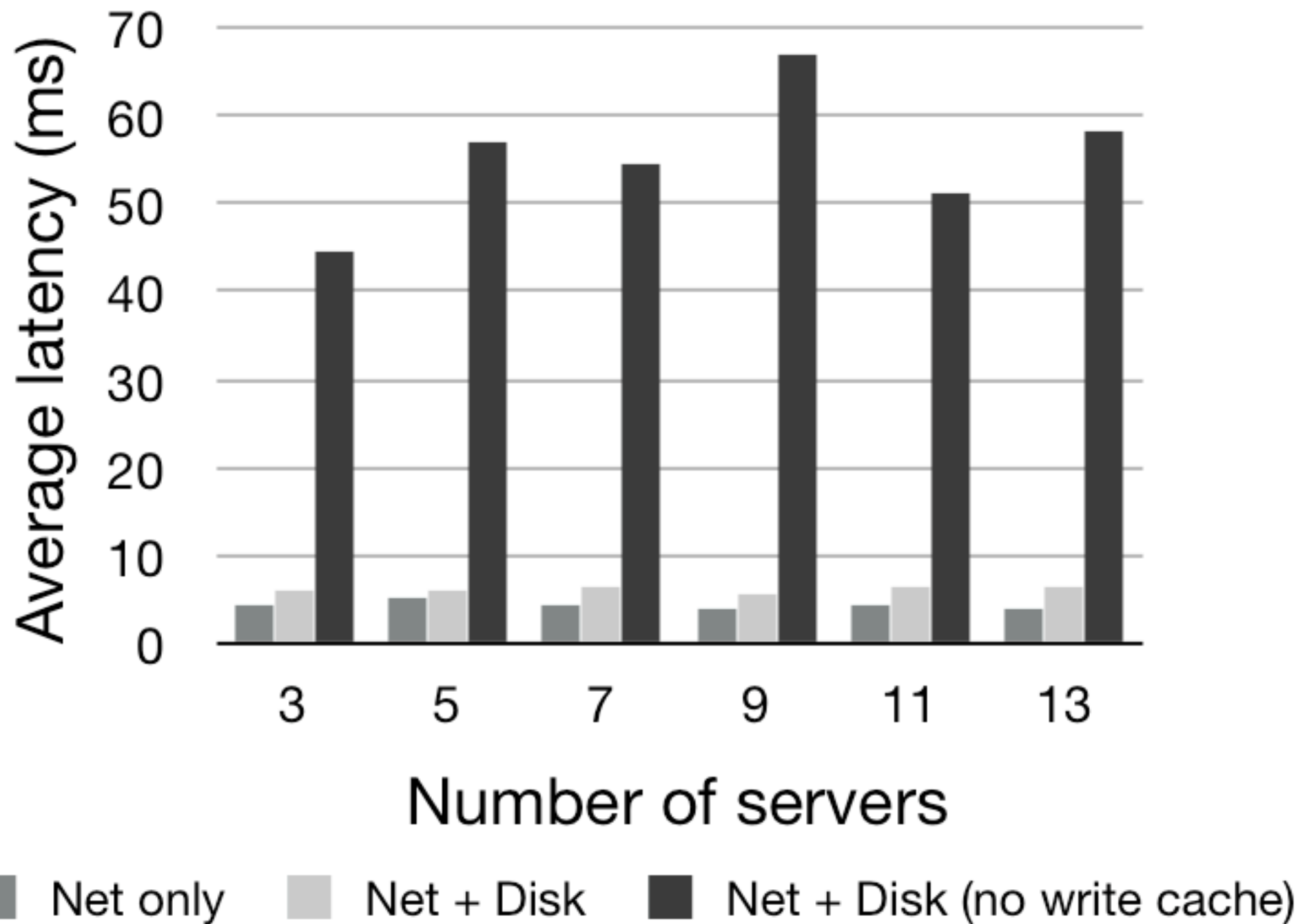
- Cluster of 50 servers
- Xeon dual-core 2.1 GHz
- 4 GB of RAM
- Two SATA disks



Throughput

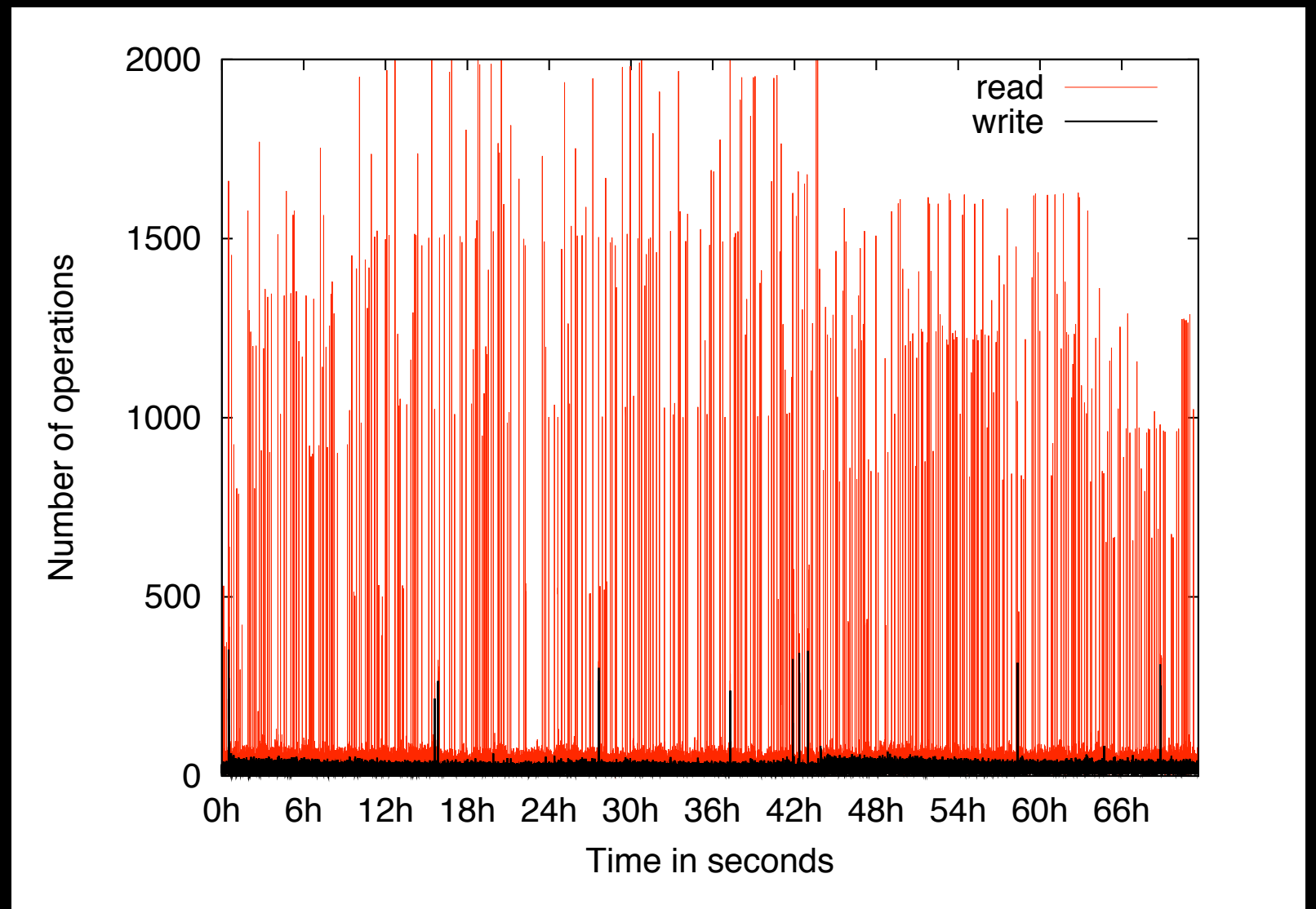


Latency



Load in production

- Fetching service
- Load of a ZKserver
- Read spikes of over 2000 reads/s
- Write spikes of less than 500 ops/s





Contributing

Steps to follow

- Watch the list for comments
- Watch issues that interest you
- Don't be shy to chip in and ask questions
- Once you're ready to contribute...

<https://cwiki.apache.org/confluence/display/ZOOKEEPER/HowToContribute>



Some important detail

- We like public communication
 - ✓ User list for questions on how to use it
 - ✓ Dev list for discussions about the code base
- Issue tracker
 - ✓ Jira <https://issues.apache.org/jira/browse/ZOOKEEPER>
 - ✓ Create a new jira issue if you:
 - ➡ Find a problem
 - ➡ Intend to contribute a patch



On our roadmap

- Dynamic configuration
 - ✓ Currently static
- Cross-colo deployments
 - ✓ POLE: Performance-Oriented Leader Election
 - ✓ Fault detection
- ZooKeeper as a Service
 - ✓ Multi-tenancy
 - ✓ Write scalability



Final remarks

- ZooKeeper
 - ✓ Service for coordination
 - ✓ Great experience internally and in Apache
- One of a few building blocks
 - ✓ BookKeeper
 - ✓ Hedwig
 - ✓ Omid





Questions?

<http://zookeeper.apache.org>