

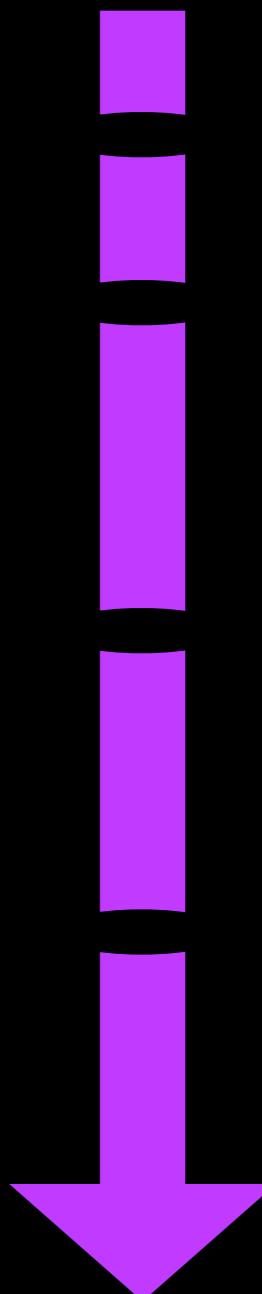


Distributed Coordination via ZooKeeper

Flavio Junqueira
Yahoo! Research, Barcelona

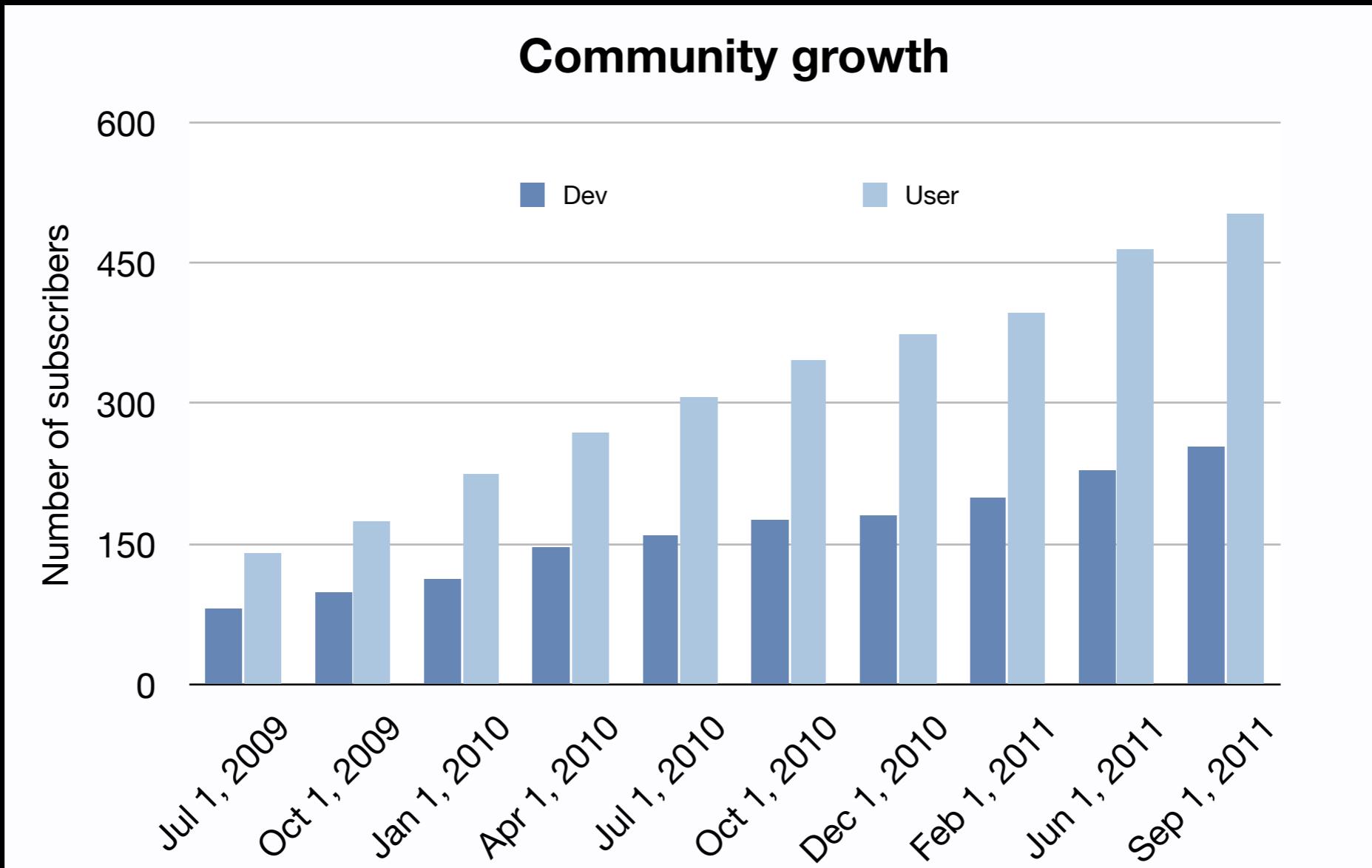
Hadoop in China 2011

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 on a variant of Paxos (it's similar to Paxos). When creating a path, delete a path, and list the children of a path or the value of a node, it's automatically removed by the ZK servers 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 Muthukaruppan, 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 on a variant of Paxos (it's similar to Paxos). When creating a path, delete a path, and list the children of a path or the value of a node, it's 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. 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 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 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.
- Scalr - Scalr uses Zookeeper to manage distributed locks. This allows Scalr to safely coordinate distributed operations across multiple hosts.

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 Zookeeper on a variant of Paxos (it's similar to Paxos, but it's faster). It's a distributed consensus algorithm that allows us to path, delete a path, and list the children of a path or the value of a node. When a node is deleted, its children are automatically removed by the ZK server. This is how Zookeeper handles consistency. 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 Mesos, and use Zookeeper for fail-over.

- ResourceManager - The ResourceManager uses Apache ZooKeeper for fail-over. When the ResourceManager fails, a secondary can quickly recover via cluster state saved in ZooKeeper.
- Scalr - Scalr uses Zookeeper to manage its distributed cron jobs. It stores the cron jobs in Zookeeper and executes them periodically. This allows Scalr to run cron jobs on multiple servers without running the risk of multiple executions or failure to execute.

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

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 You

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 * 70° / 49° 61° / 42° 63° / 47°

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

* Severe weather alert

City or ZIP Go

Yahoo! Notícias: Foto de Portada Options

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

Yahoo Mail Preview Options

fjunqueira Lufthansa 03:37 am

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

word@m-w.com Oct 24

grunlette: 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 My Yahoo! portal homepage with several features highlighted by callout boxes:

- Search**: Points to the search bar at the top.
- E-mail**: Points to the Mail icon in the Personal Assistant sidebar.
- Finance**: Points to the Stocks icon in the Personal Assistant sidebar.
- Weather**: Points to the weather forecast for Philadelphia, PA.
- News**: Points to the Yahoo! News: Most Viewed section.

Personal Assistant sidebar:

- Mail
- Horoscope
- Stocks
- Weather 57°F
- Lottery
- Sports

Message Center sidebar:

Weather sidebar:

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

Yahoo! News: Foto de Portada section:

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

Yahoo! Mail Preview section:

fjunqueira

- Lufthansa: Travel information for your flight on 28/Oct to Frankfurt/Main, Mr Junqueira (03:37 am)
- word@m-w.com: word@m-w.com (Oct 24)
- grunlette: M-W's Word of the Day (Oct 23)
- Ryan Schmidt: Barra in the Times (Oct 23)

Travelocity Deals: Save up to 40%, plus affordable Asia vacations (Oct 23)

Economist.com: New on Economist.com - 23rd October 2009 (Oct 23)

Compose | Refresh

Yahoo! News: Most Viewed section:

- 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!: Workload generated



Yahoo!: Workload generated

Web Images Video Local Apps More ▾

YAHOO!

Y! Schweiz | My Yahoo! | Make Y! your homepage

Hi, Flavio | Sign Out | Page Options

YAHOO SITES Edit

- Mail
- Autos
- Dating
- Finance (Dow Jones)
- Flickr
- Games
- Health
- Horoscopes
- Jobs
- Messenger
- Movies
- News
- omg!
- Real Estate
- Shine
- Shopping
- Sports
- Travel
- TV
- Weather (72°F)

More Yahoo Sites

MY FAVORITES Edit

- Facebook

TODAY - October 09, 2011



Things that claim to work, but don't

Does the walk signal really change because you press the button, or is it a placebo? The truth behind it »

- What is placebo?
- Unhealthy health drinks
- Do toning shoes work?

5 things that don't work | Paris Jackson dresses like dad | Classy gesture after MLB win | Glamour guys in big NFL game

1 - 4 of 24

NEWS WORLD LOCAL FINANCE

- 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

| MLB | NFL | College football | NASCAR | Soccer | UFC

updated 05:51 pm More: News Popular Photos

TRENDING NOW

- 1. Nancy Shevell
- 2. Michelle Williams
- 3. Ben Stiller
- 4. Casey Anthony
- 5. Consumer credit
- 6. Evangeline Lilly
- 7. UFC 136
- 8. Sesame Street
- 9. H1N1 treatment
- 10. Home ownership

AdChoices

Yahoo! Search Search Web: Go

Anne Hathaway Jennifer Lopez Will Smith House American Idol Jersey Shore Breakfast Recipes Pancakes French Toast

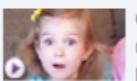
VIDEO PICKS



Go to Video

Haunted house camera captures faces of fear

 Vacuum gives kittens a run for their lives

 Girl's priceless reaction to birthday surprise



Yahoo!: Workload generated

YAHOO!®

Web Images Video Local Apps More

Y! Schweiz | My Yahoo! | Make Y! your homepage

Hi, Flavio | Sign Out | Page Options

YAHOO SITES Edit

TODAY - October 09, 2011

Yahoo! Homepage: 4.4 billion page views a month

YAHOO SITES

- Mail
- Horoscopes
- Jobs
- Messenger
- Movies
- News
- omg!
- Real Estate
- Shine
- Shopping
- Sports
- Travel
- TV
- Weather (72°F)

More Yahoo Sites

MY FAVORITES Edit

- Facebook

updated 05:51 pm

More: News Popular Photos

Web Search

TRENDING NOW

- 1. Nancy Shevell
- 6. Evangeline Lilly
- 2. UFC 136
- 7. Sesame Street
- 3. H1N1 treatment
- 8. Home ownership
- 4. Williams
- 9. Tony
- 5. Credit

AdChoices

Search Web: Go

Does the walk signal really change because you press the button, or is it a placebo? The truth behind it »

• What is a placebo?
• Unhealthy health drinks
• Do toning shoes work?

5 things that don't work

Paris Jackson dresses like dad

Classy gesture after MLB win

Glamour guys in big NFL game

1 - 4 of 24

Anne Hathaway Jennifer Lopez Will Smith House American Idol Jersey Shore Breakfast Recipes Pancakes French Toast

NEWS WORLD LOCAL FINANCE

- 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

| MLB | NFL | College football | NASCAR | Soccer | UFC

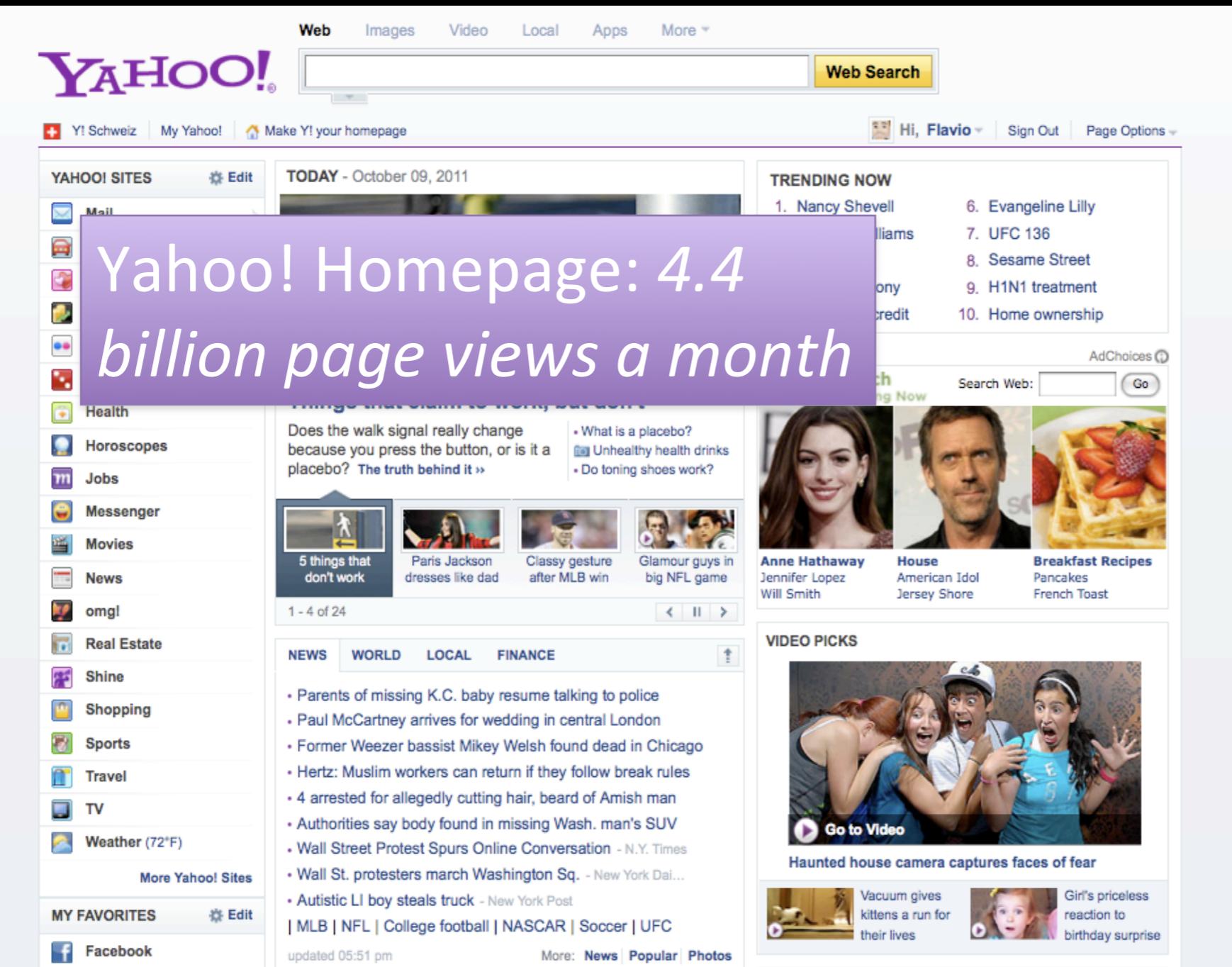
VIDEO PICKS

Haunted house camera captures faces of fear

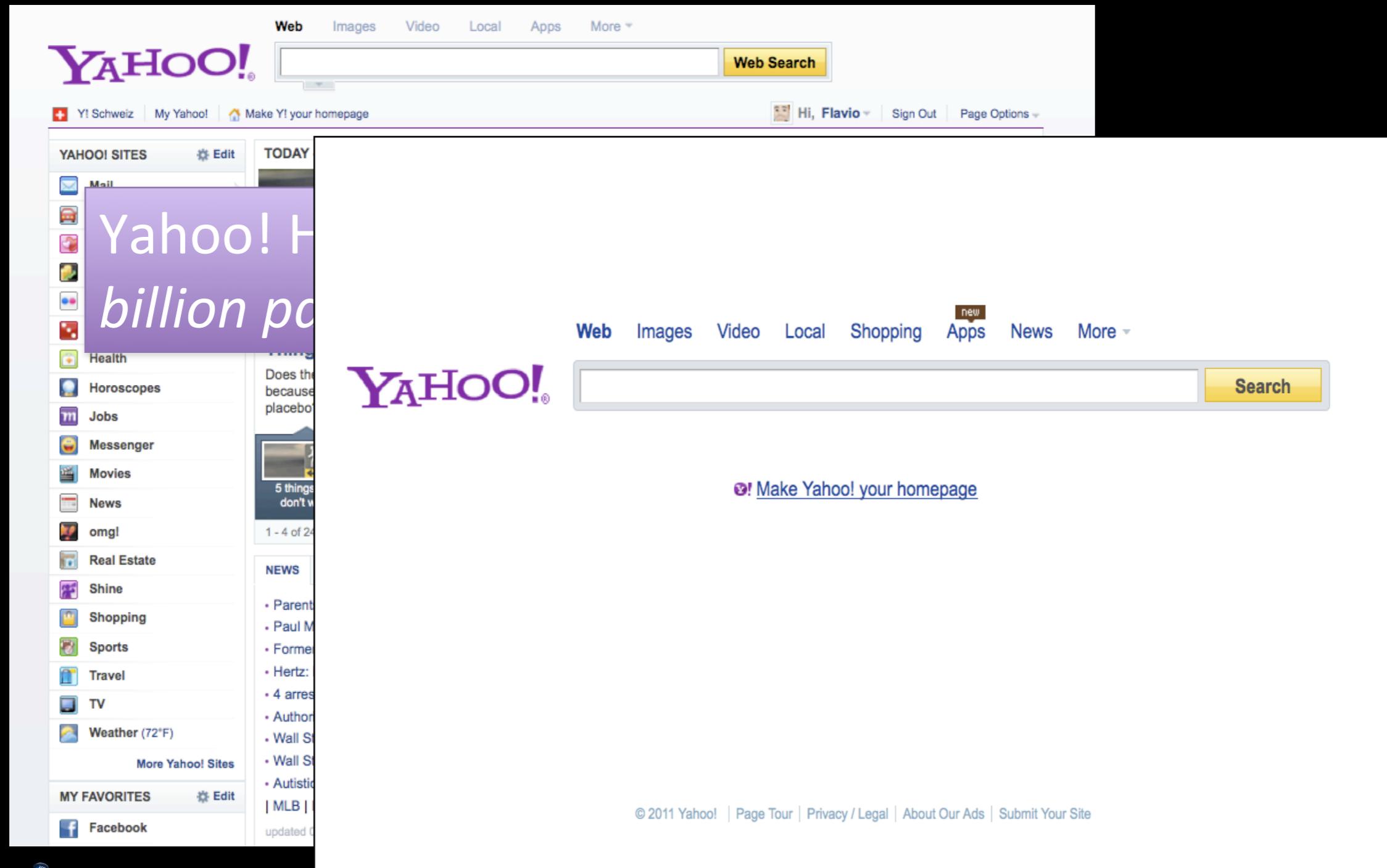
Go to Video

Vacuum gives kittens a run for their lives

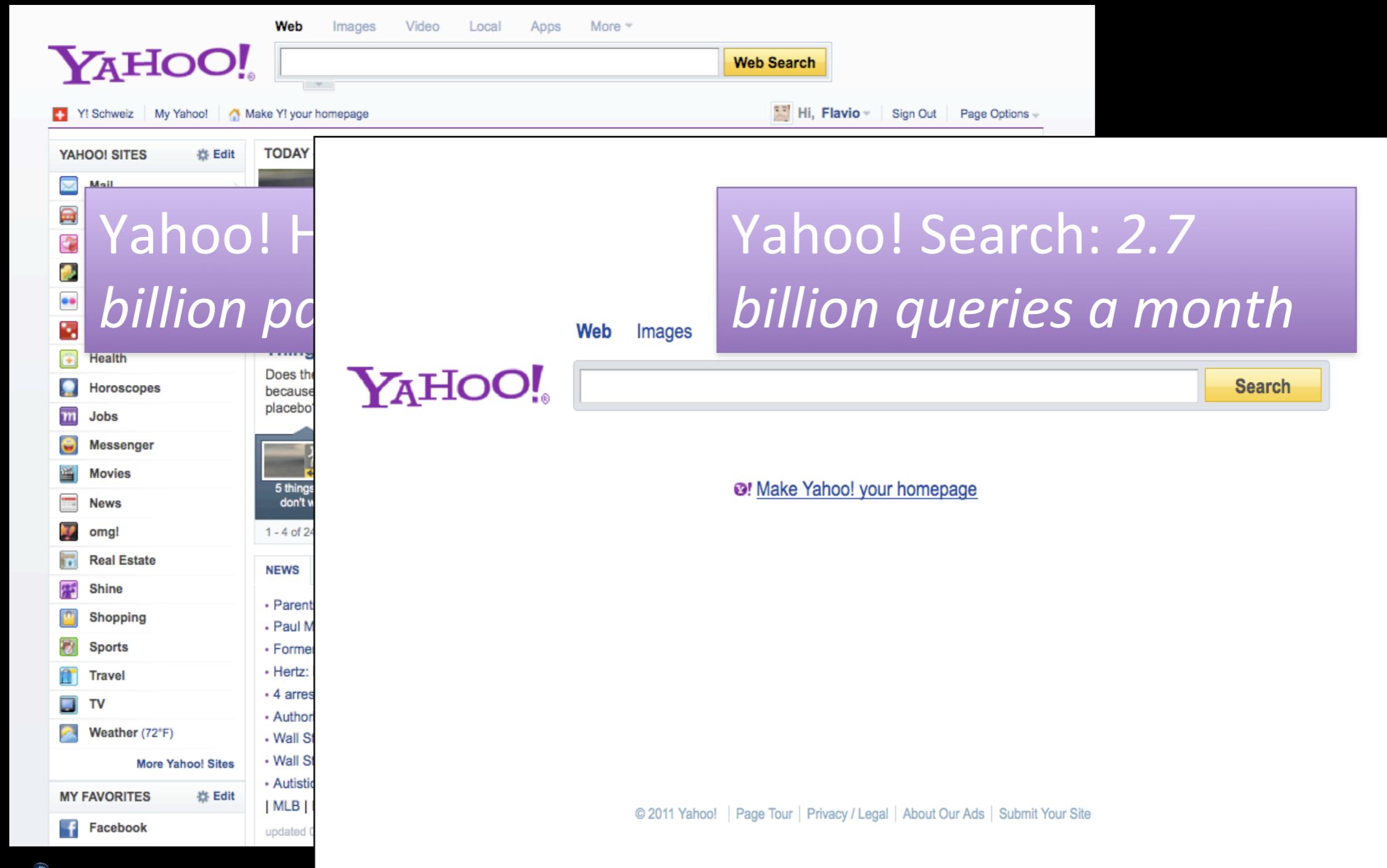
Girl's priceless reaction to birthday surprise



Yahoo!: Workload generated



Yahoo!: Workload generated



Yahoo!: Workload generated

Web Images Video Local Apps More

Hi, Flavio | Sign Out | Help

YAHOO! NEWS

Make Y! home, help a school

Search Web

Search

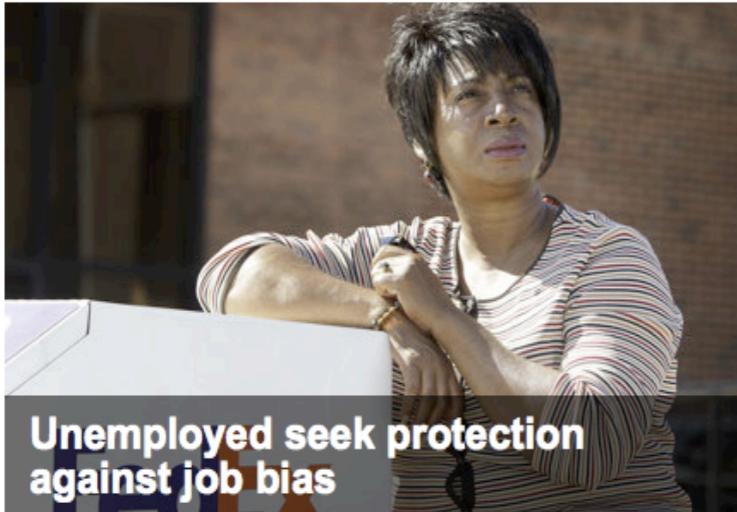
HOME U.S. WORLD BUSINESS ENTERTAINMENT SPORTS TECH POLITICS SCIENCE HEALTH BLOGS LOCAL POPULAR

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


Immigrants fearing deportation make plans for kids


Anti-Gadhafi fighters make gains in Sirte

MOST POPULAR »

- 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
- Ron Paul Wins Straw Poll; Cain 2nd

AdChoices ▾

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 face their financial fears

Top Stories Latest News Slideshows AP Reuters AFP

 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 »](#)

in: 2.7
s a month



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 that is the Yahoo! logo and a search bar with the placeholder "Search". To the right of the search bar are links for Mail, My Y!, and a home icon. The main header says "YAHOO! NEWS". Below the header is a menu bar with categories: HOME, U.S., WORLD, BUSINESS, ENTERTAINMENT, SPORTS, TECH, POLITICS, SCIENCE, HEALTH, BLOGS, LOCAL, and POPULAR. Underneath this are more links for VIDEO, PHOTOS, GMA, LOCAL, ODD NEWS, COMICS, TRAVEL, OPINION, TRENDING NOW, VITALITY, WANDERLUST, WHO KNEW?, and WEATHER. A prominent purple banner in the center-left reads: "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." It includes a call-to-action button "To get started, first Login with Facebook". Below this banner is a large image of a person's face. To the right of the banner, there's a sidebar with a list of news items. Further down, there are two main news stories: "Unemployed seek protection against job bias" and "Anti-Gadhafi fighters make gains in Sirte". At the bottom, there are buttons for Top Stories, Latest News, Slideshows, AP, Reuters, and AFP. A small image of a crowd is on the left of the stories.

YAHOO! NEWS

Make Y! home, help a school

Hi, Flavio | Sign Out | Help

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](#)

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

Unemployed seek protection against job bias

Anti-Gadhafi fighters make gains in Sirte

Top Stories Latest News Slideshows AP Reuters AFP

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 »](#)

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 face their financial fears

in: 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



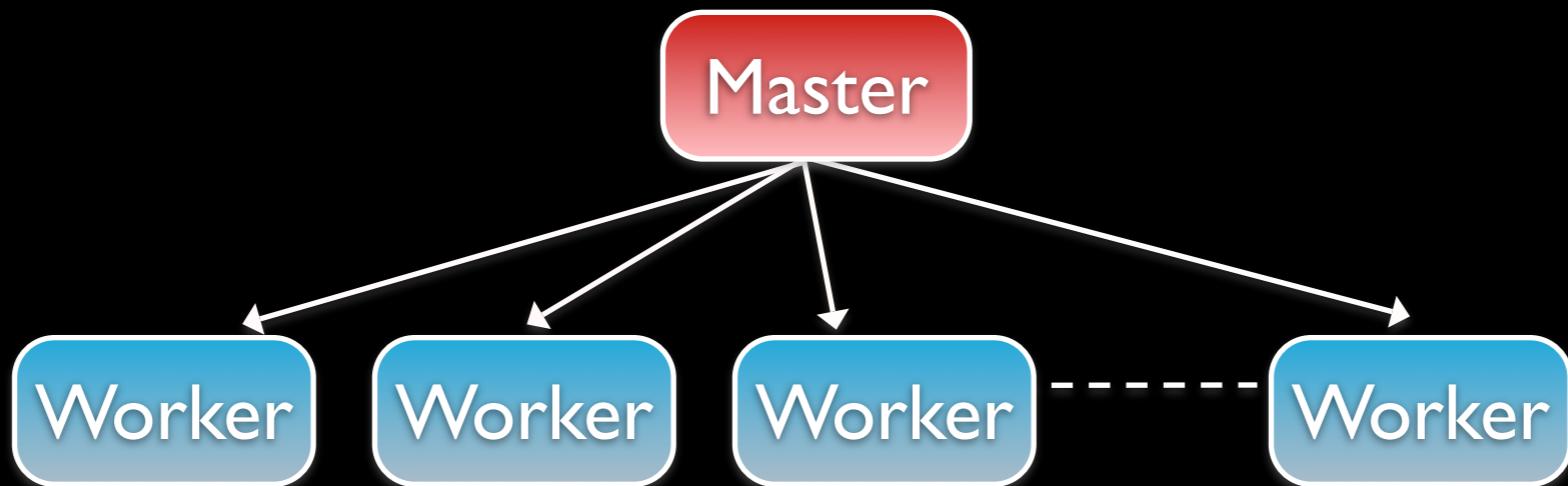
... and in computer systems?

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

Require knowledge
of complex protocols

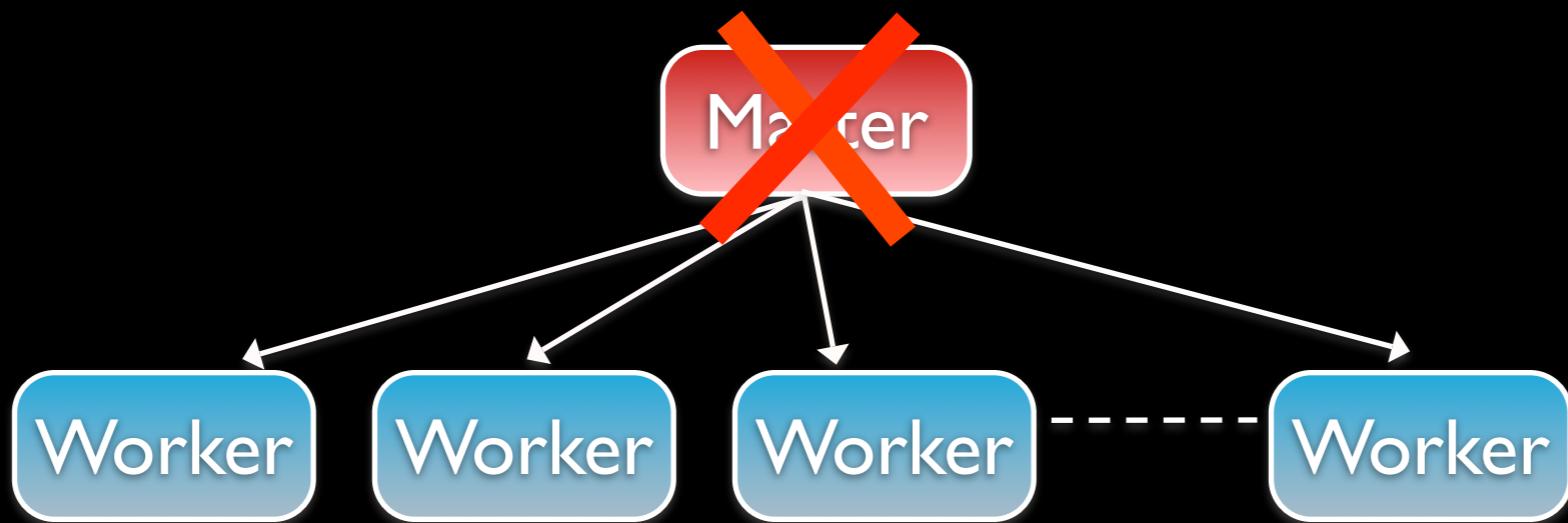
A simple model

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



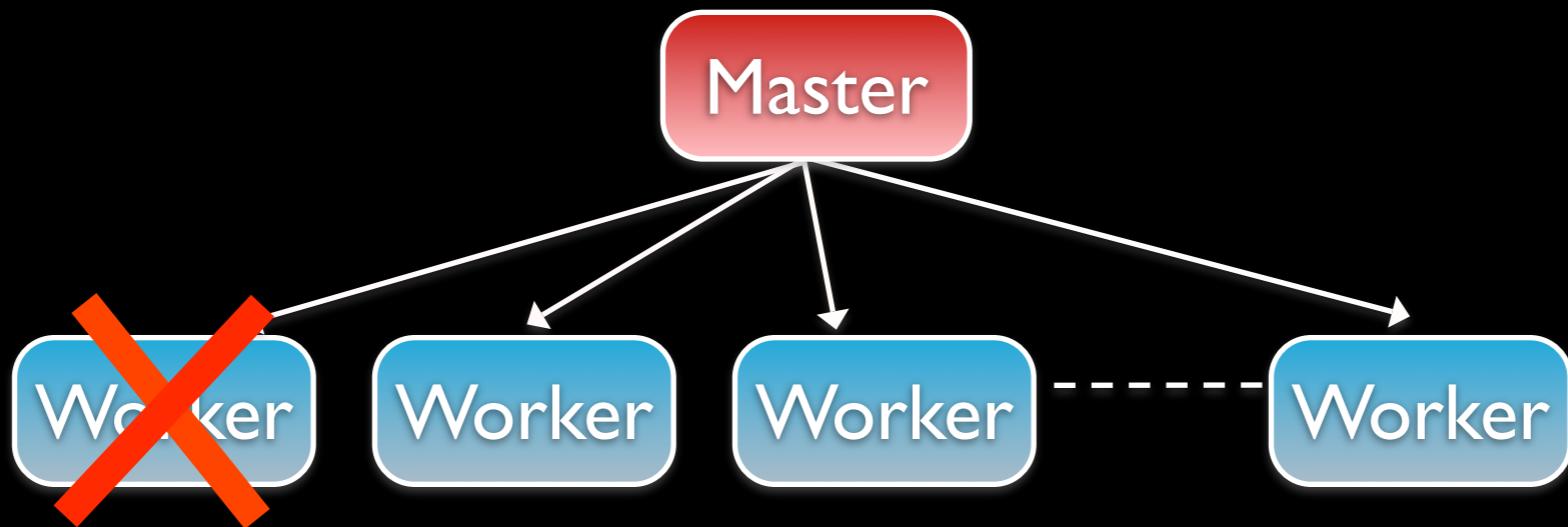
Master crashes

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



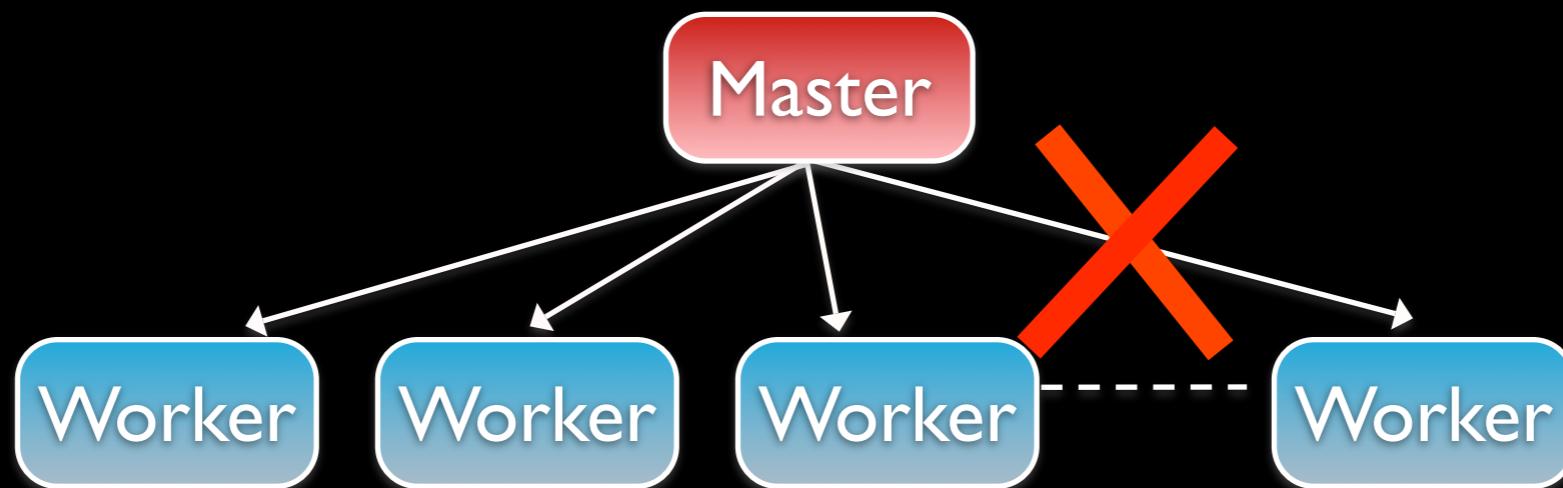
Worker crashes

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



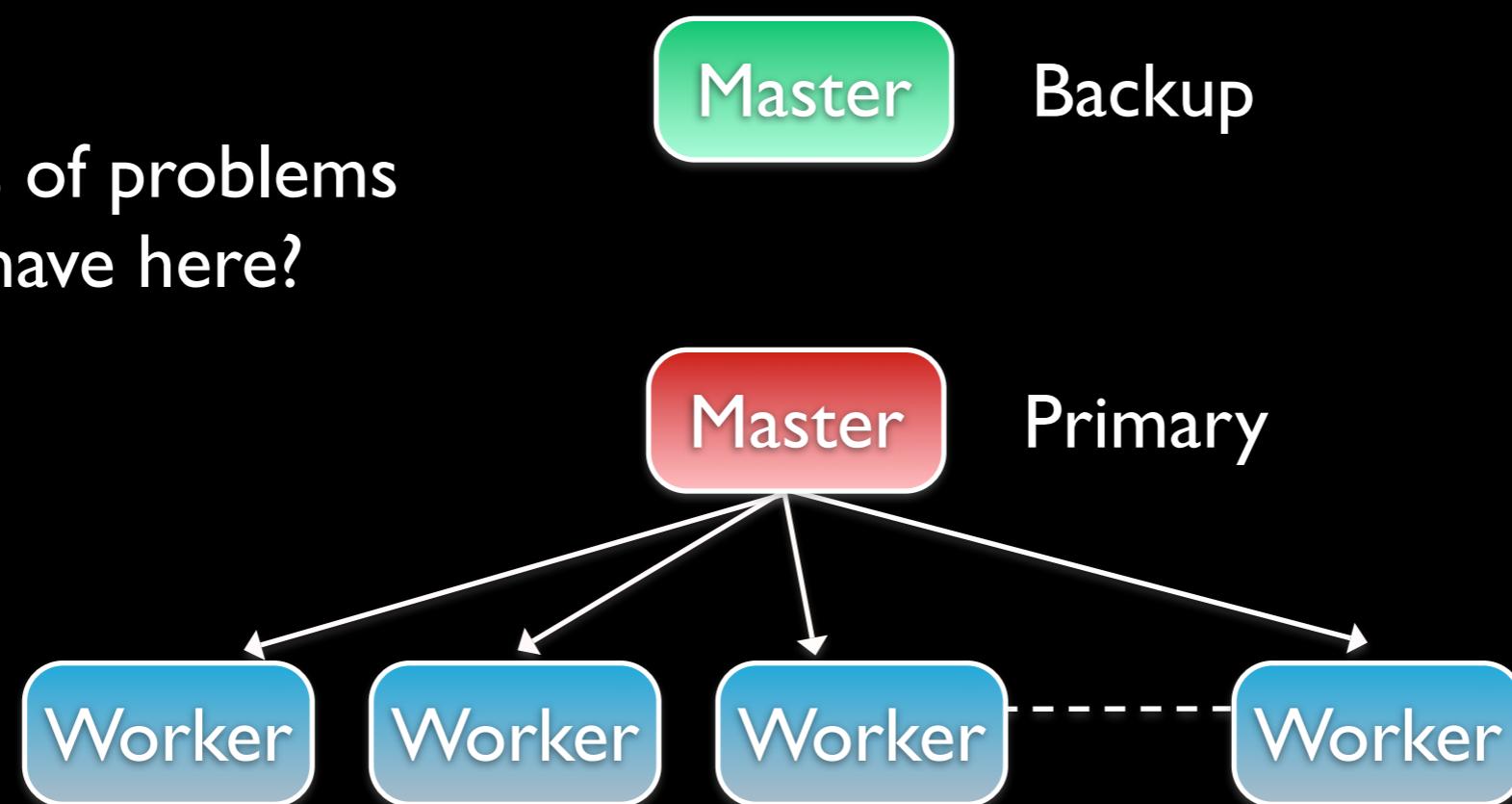
Worker does receive assignment

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

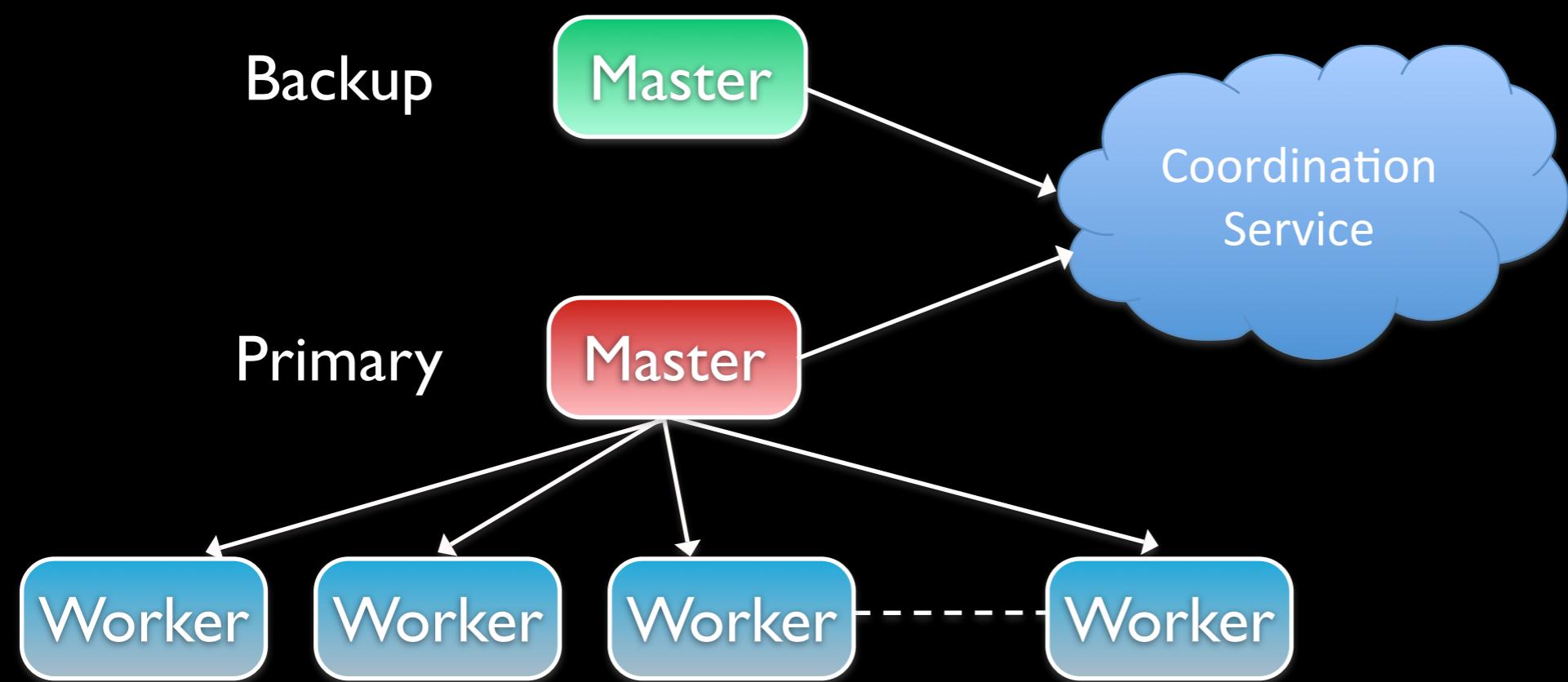


Fault-tolerant distributed system

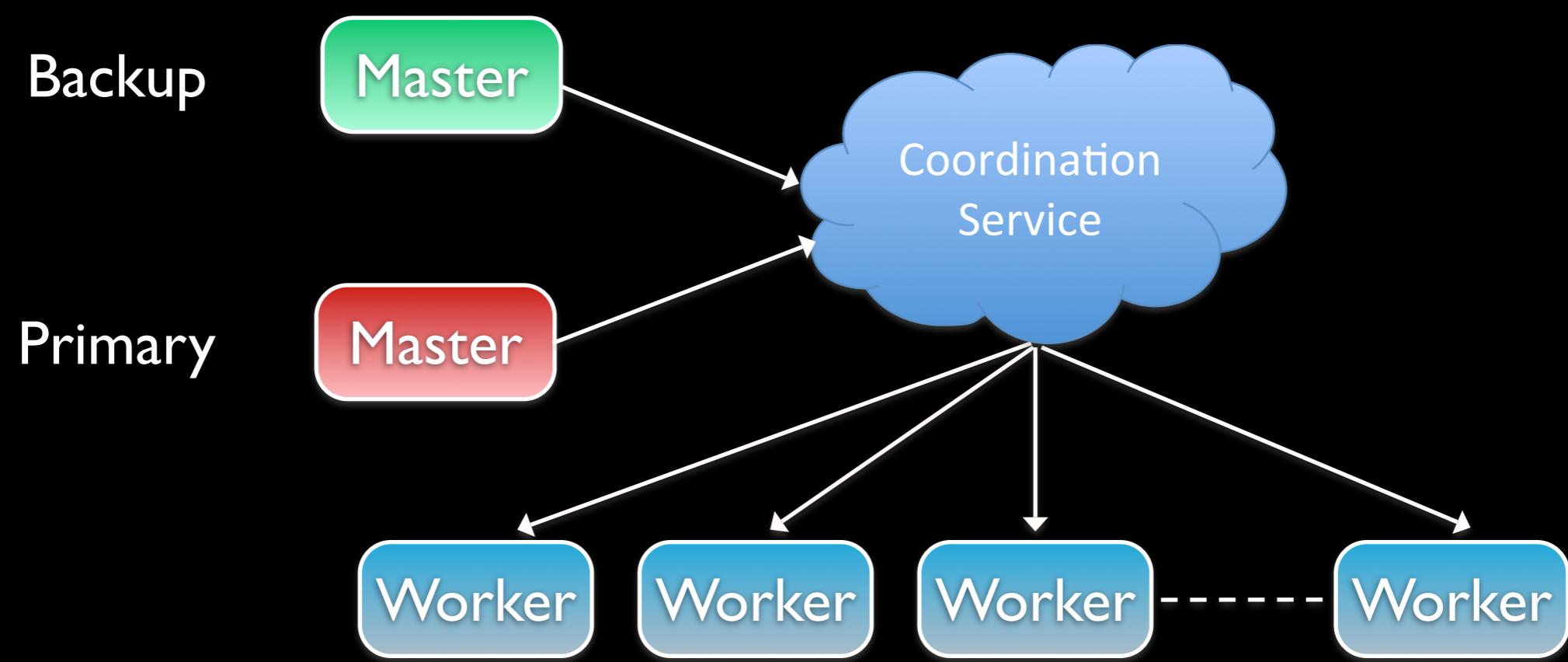
What kinds of problems
can we have here?



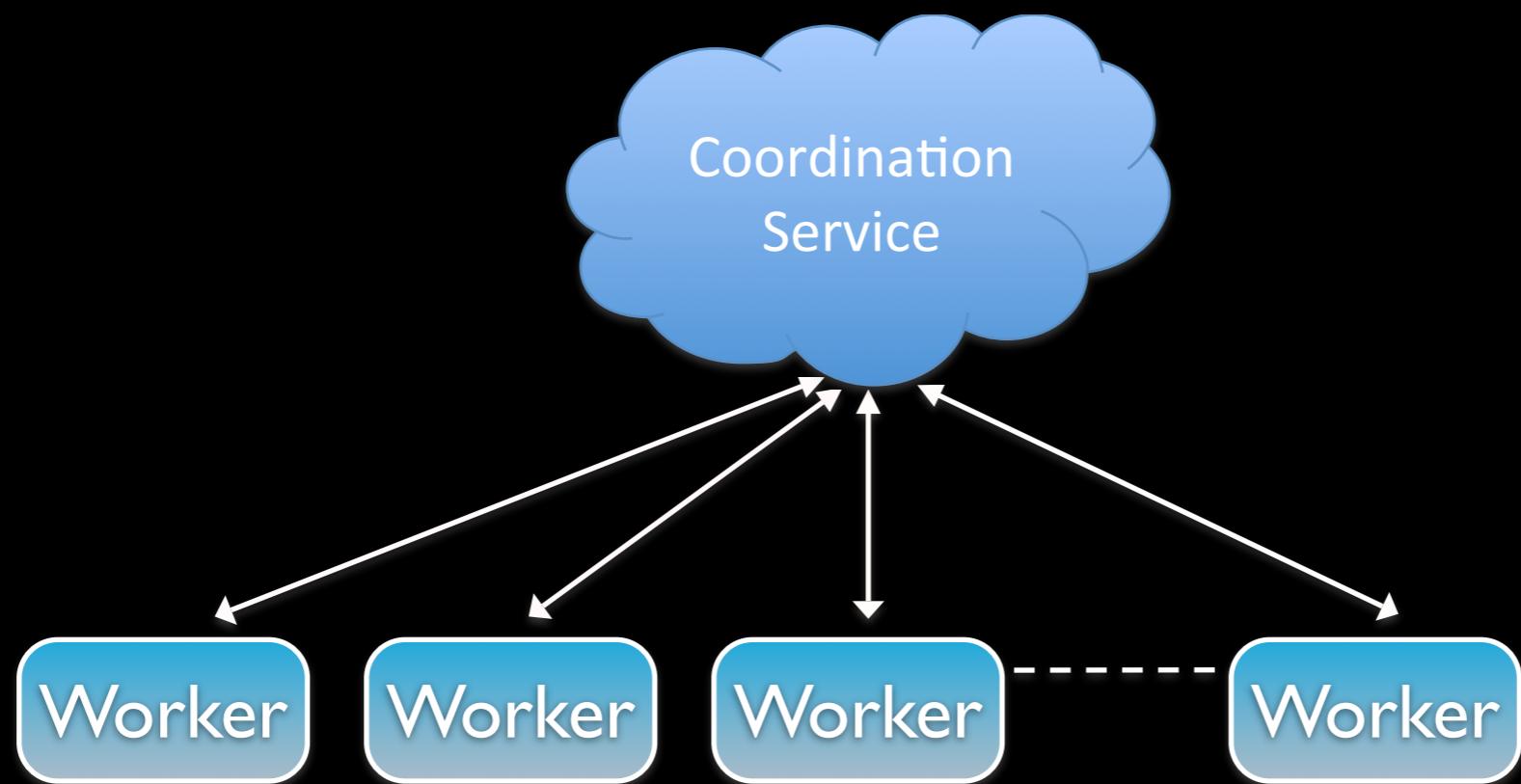
Fault-tolerant distributed system



Fault-tolerant distributed system



Fully distributed



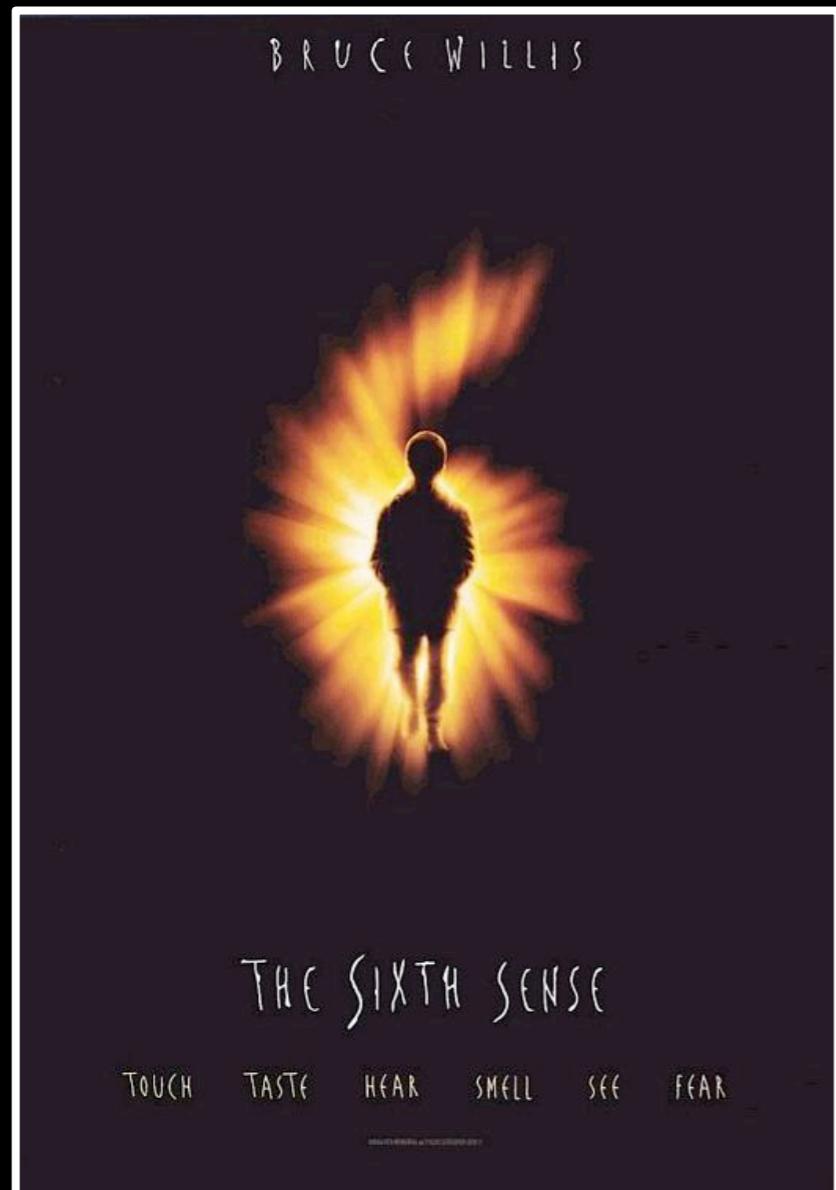
Fallacies of distributed computing

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

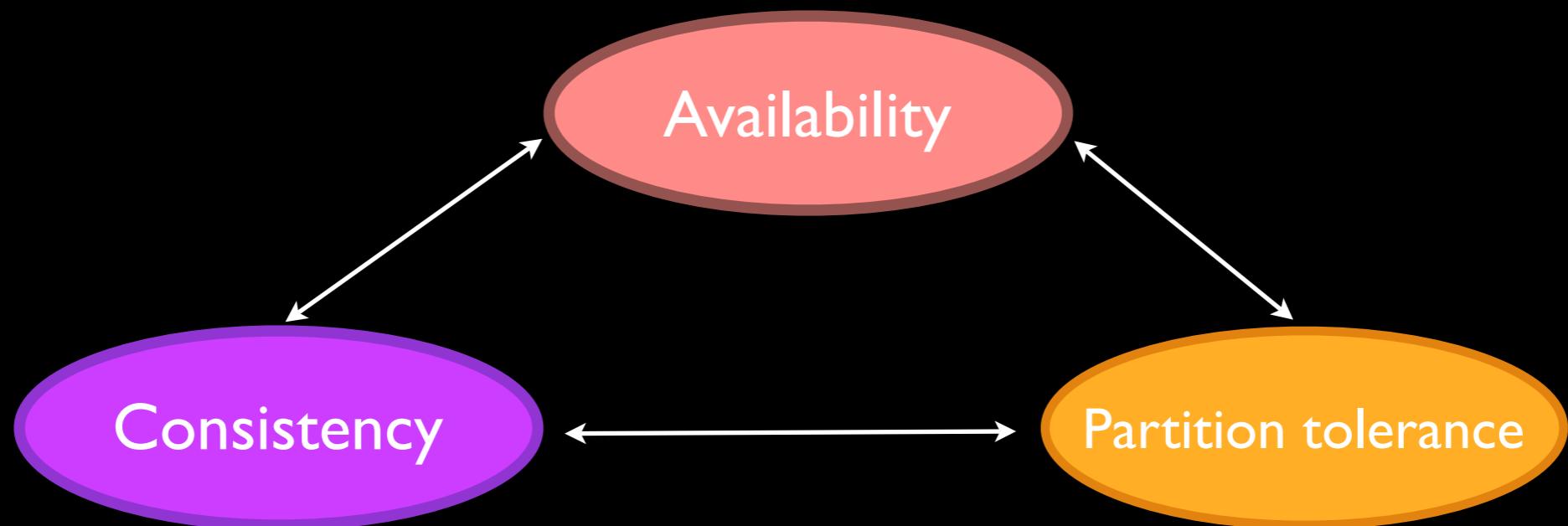
- According to Herlihy we do need consensus
 - ✓ Wait-free synchronization
 - ✓ Wait-free: Operations complete in a finite number of steps
 - ✓ Equivalent to solving consensus for n processes

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

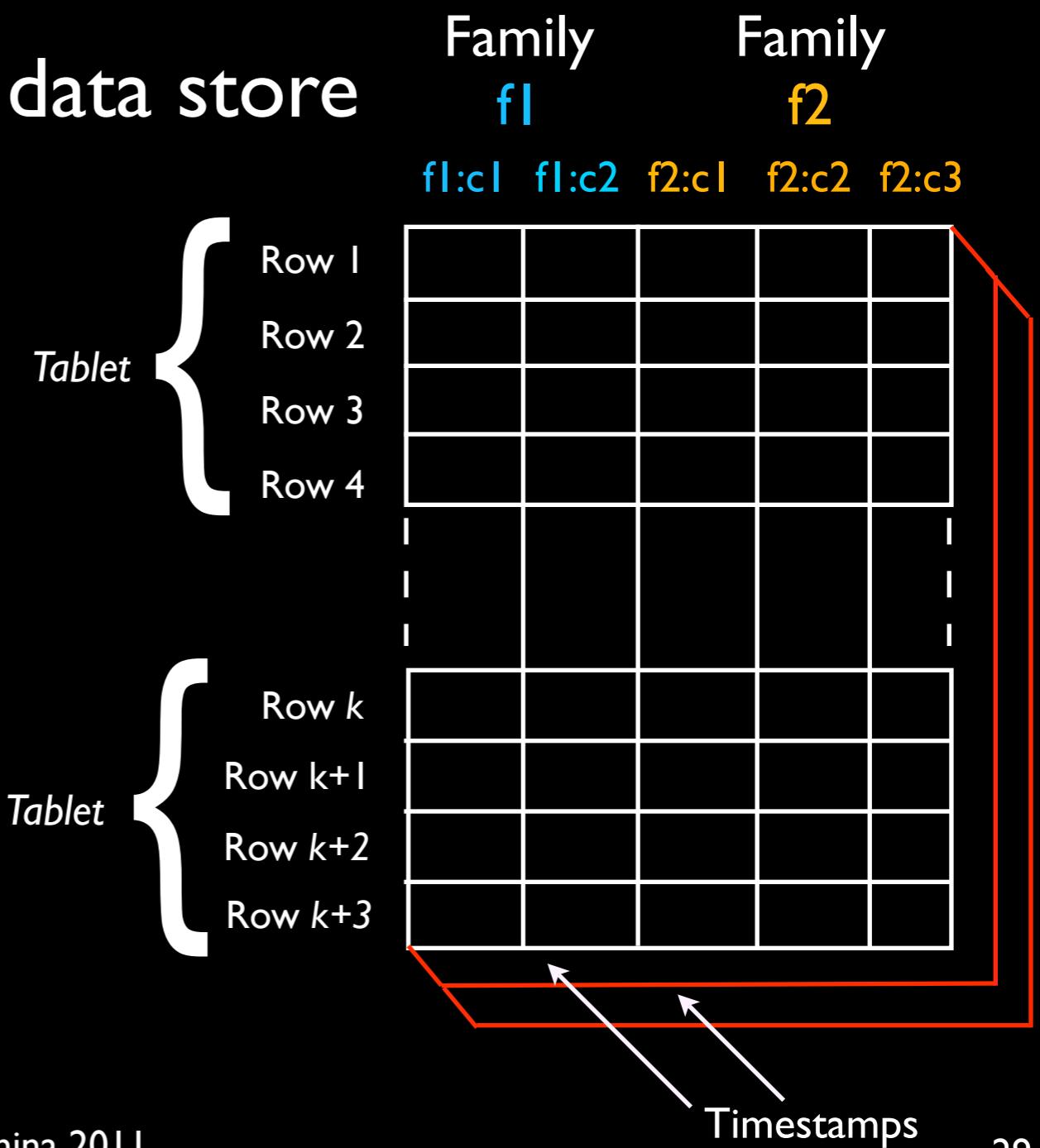
- ✓ Initially contributed by Yahoo!

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



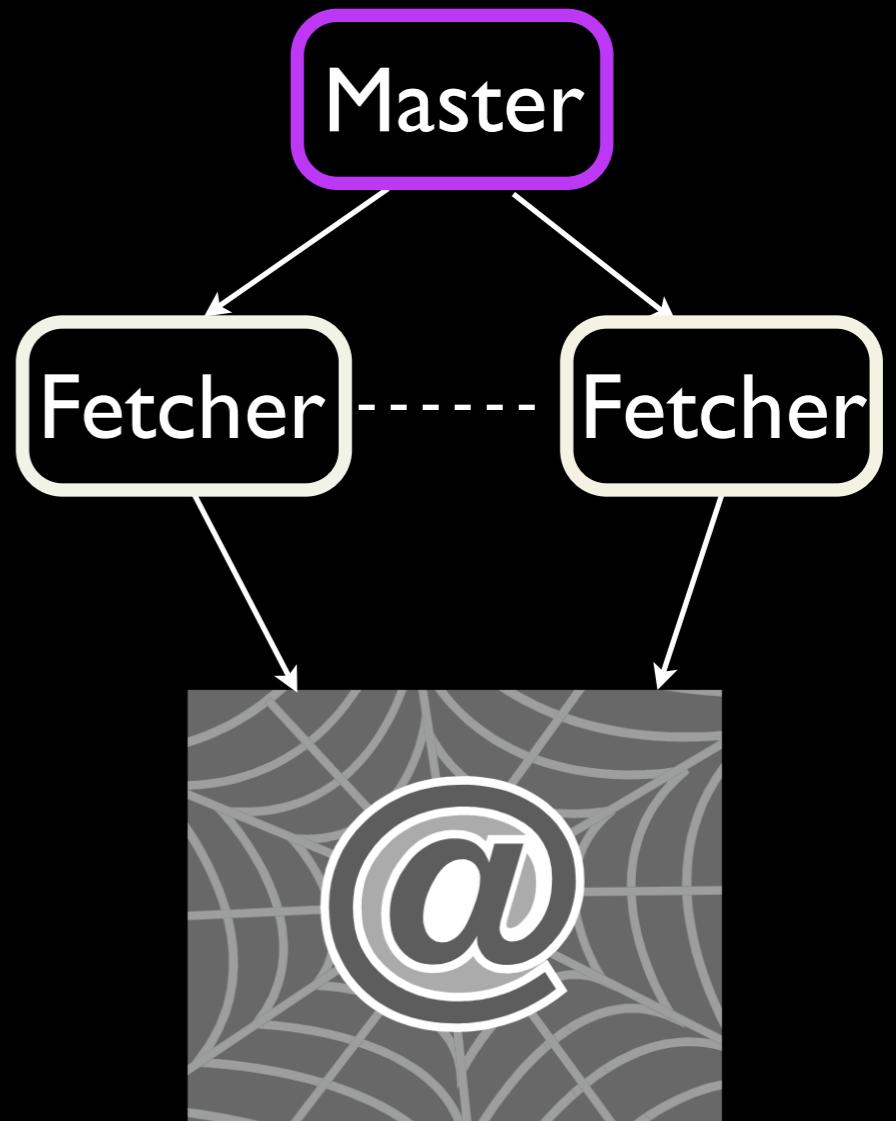
Example - Bigtable, Hbase

- 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



ZooKeeper

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

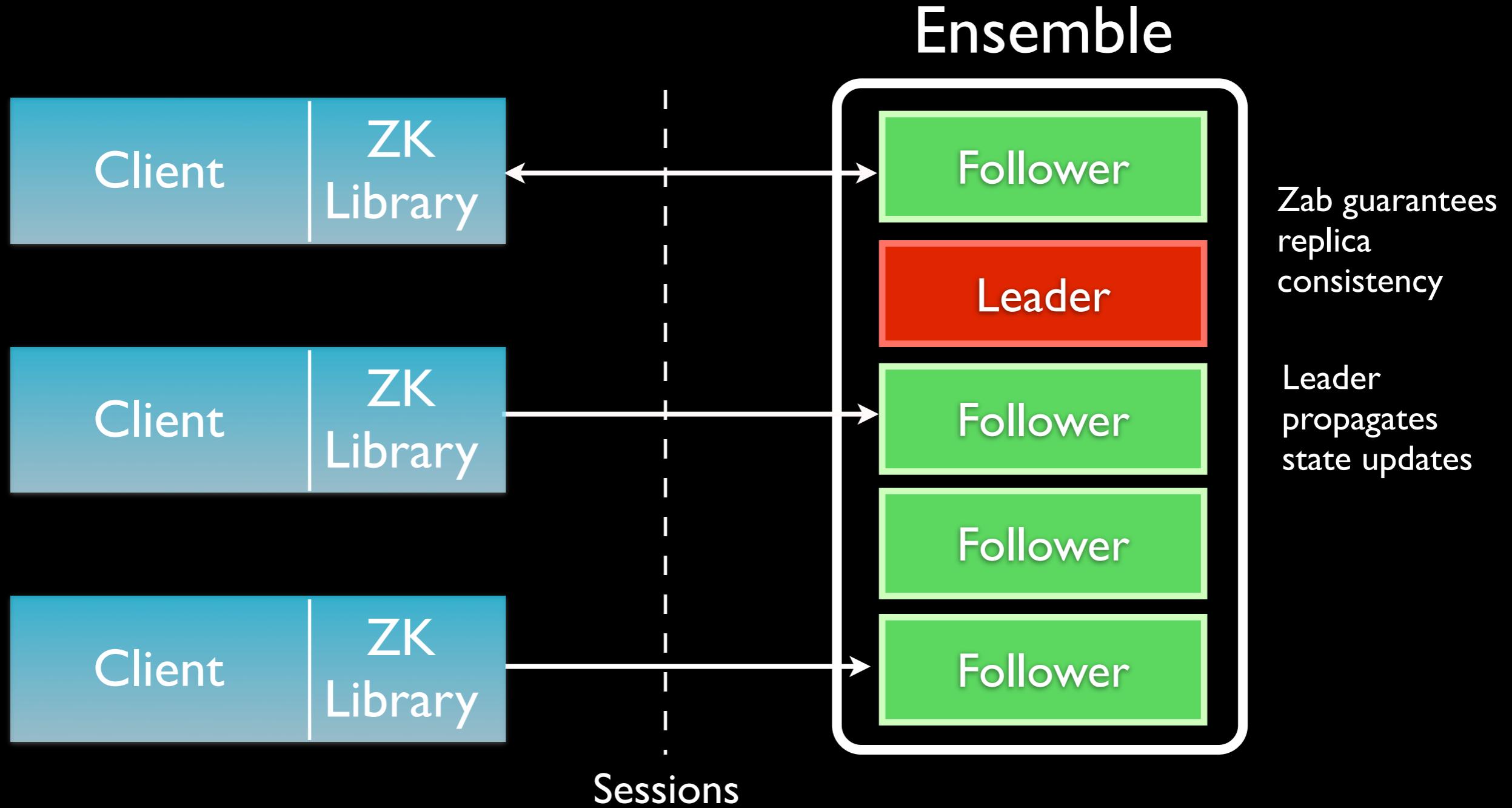


More introduction

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

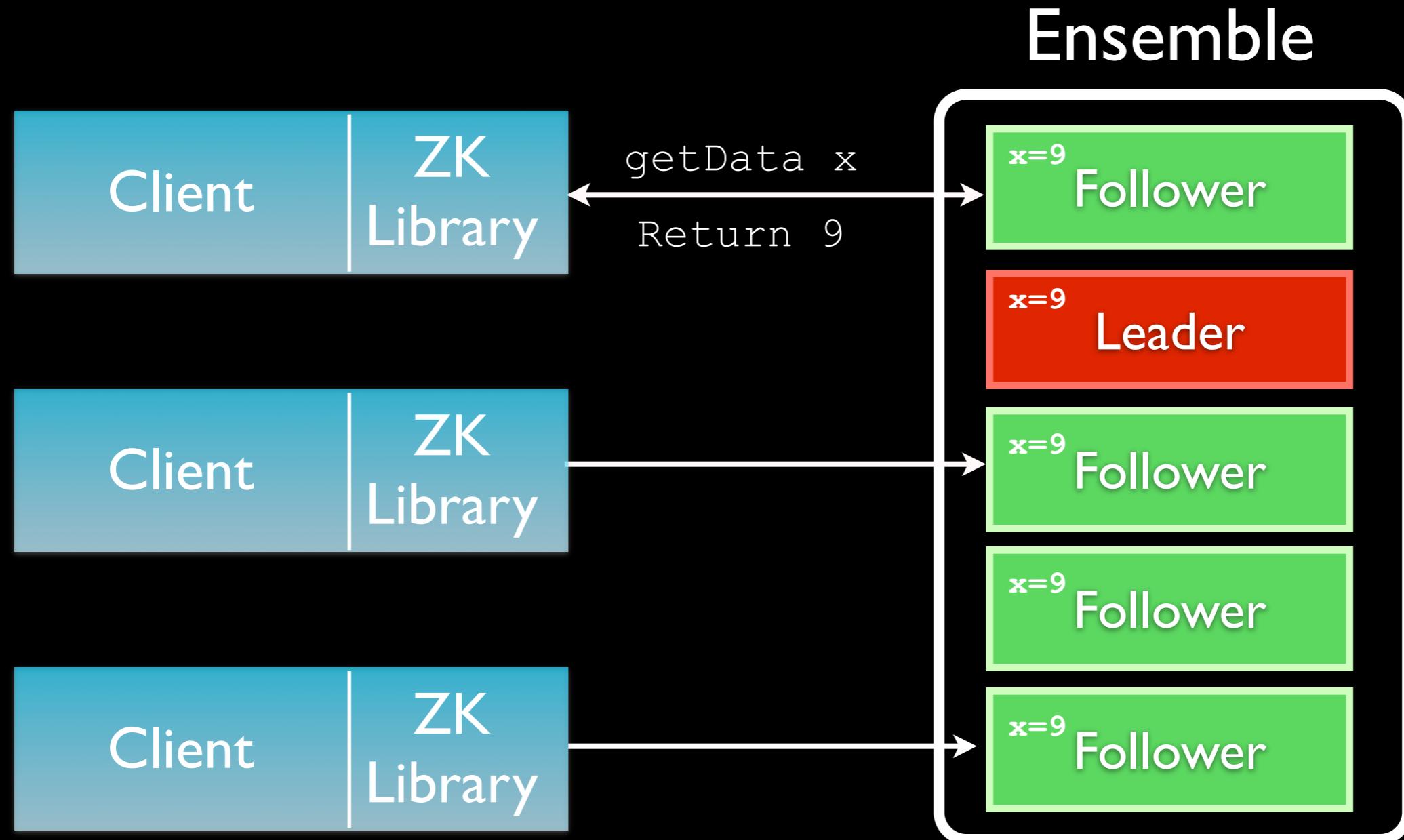


ZooKeeper API

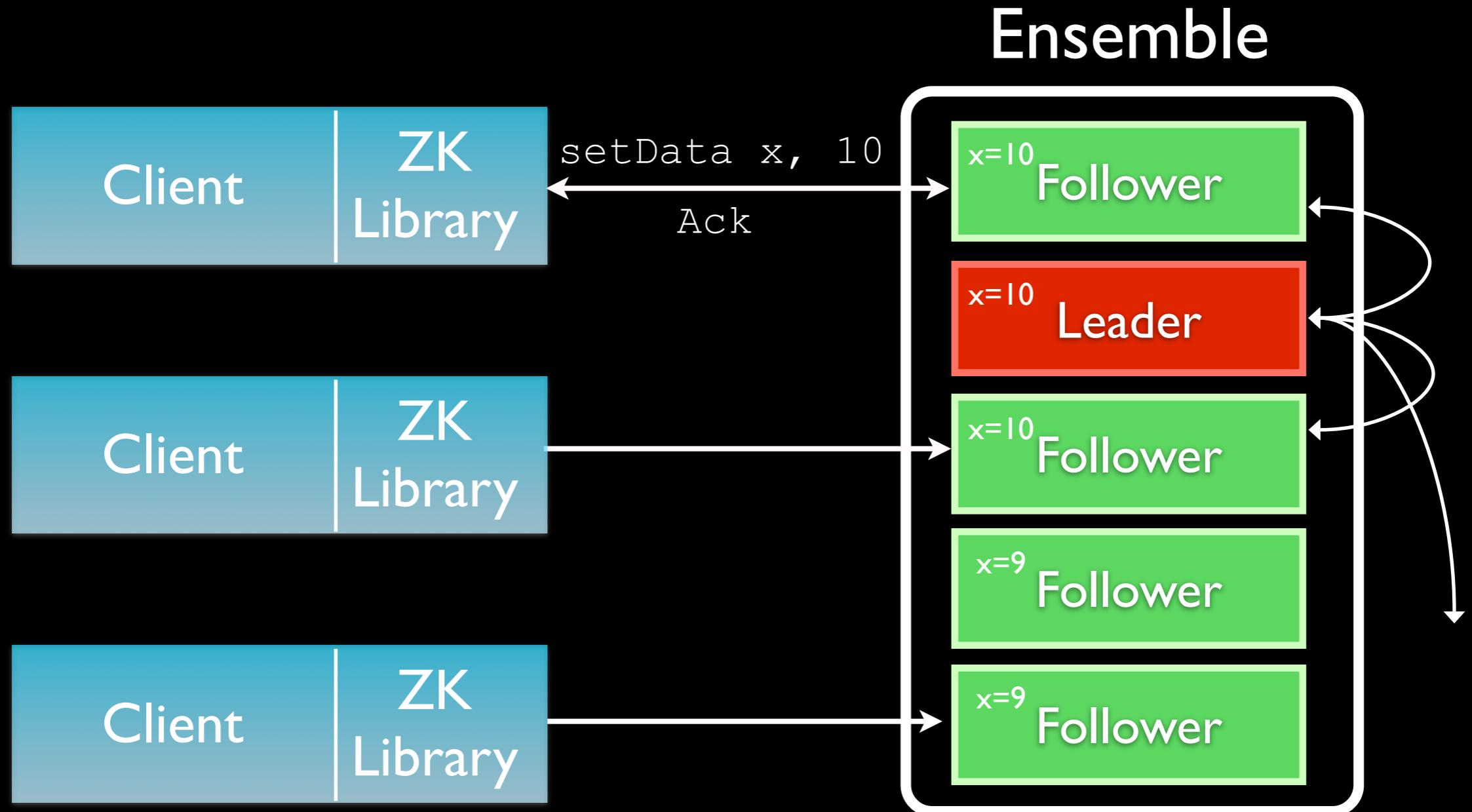
- **Create znodes:** create
 - ✓ Persistent, ephemeral, sequential
- **Read and modify data:** getData, setData
- **Read the children of znode:** getChildren
- **Check if znode exists:** exists
- **Delete znode:** delete



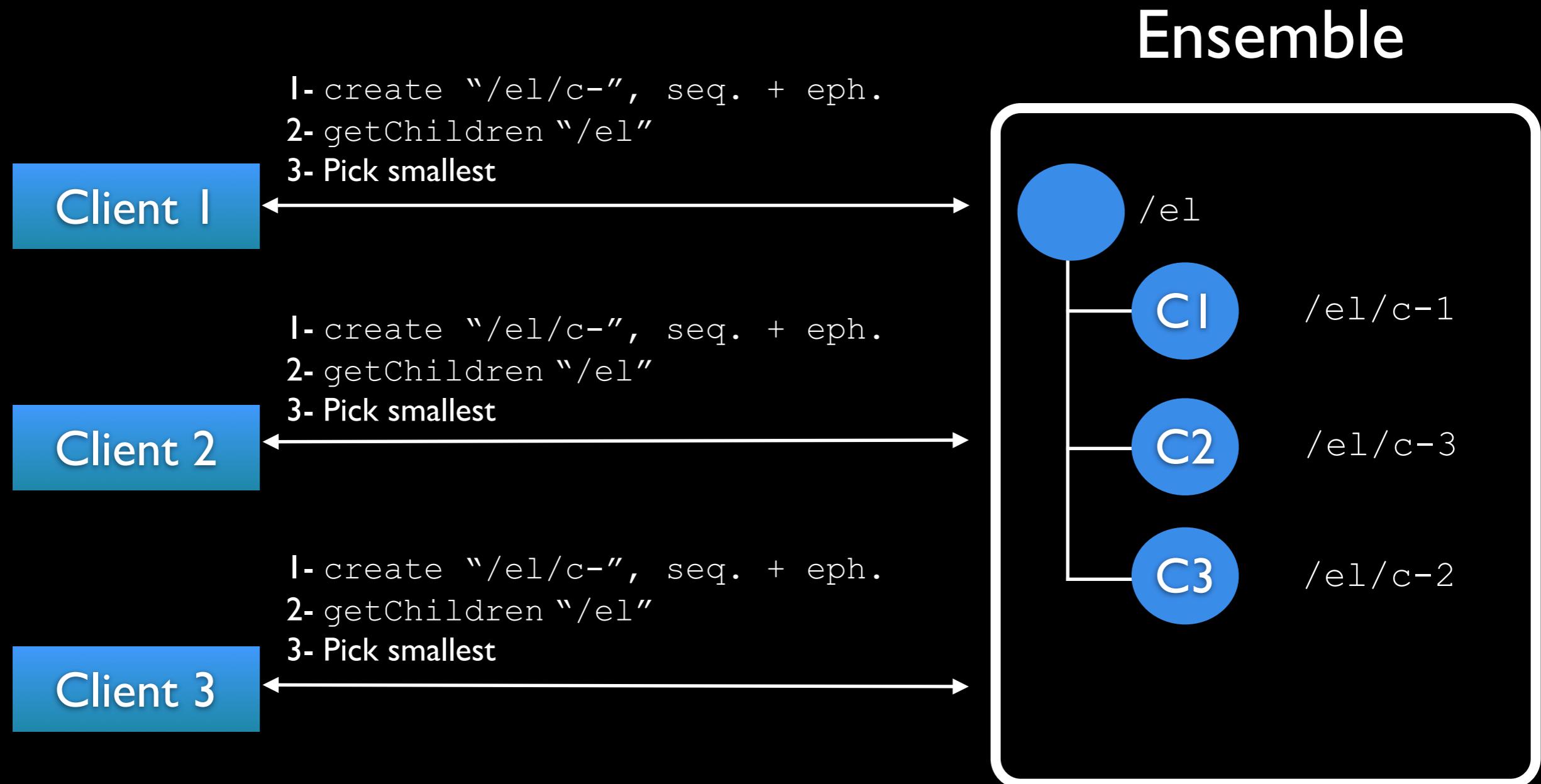
ZooKeeper: Reads



ZooKeeper: Writes



Example

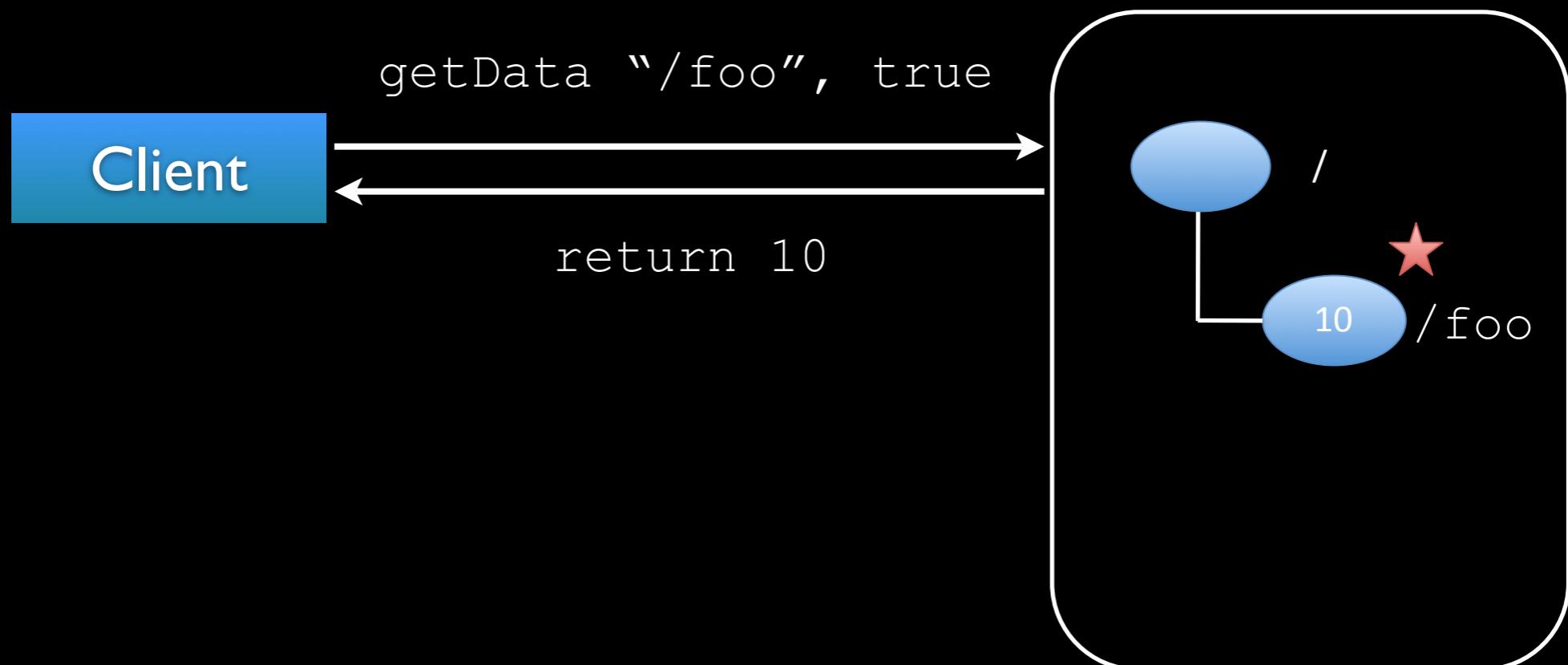


Znode changes

- 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



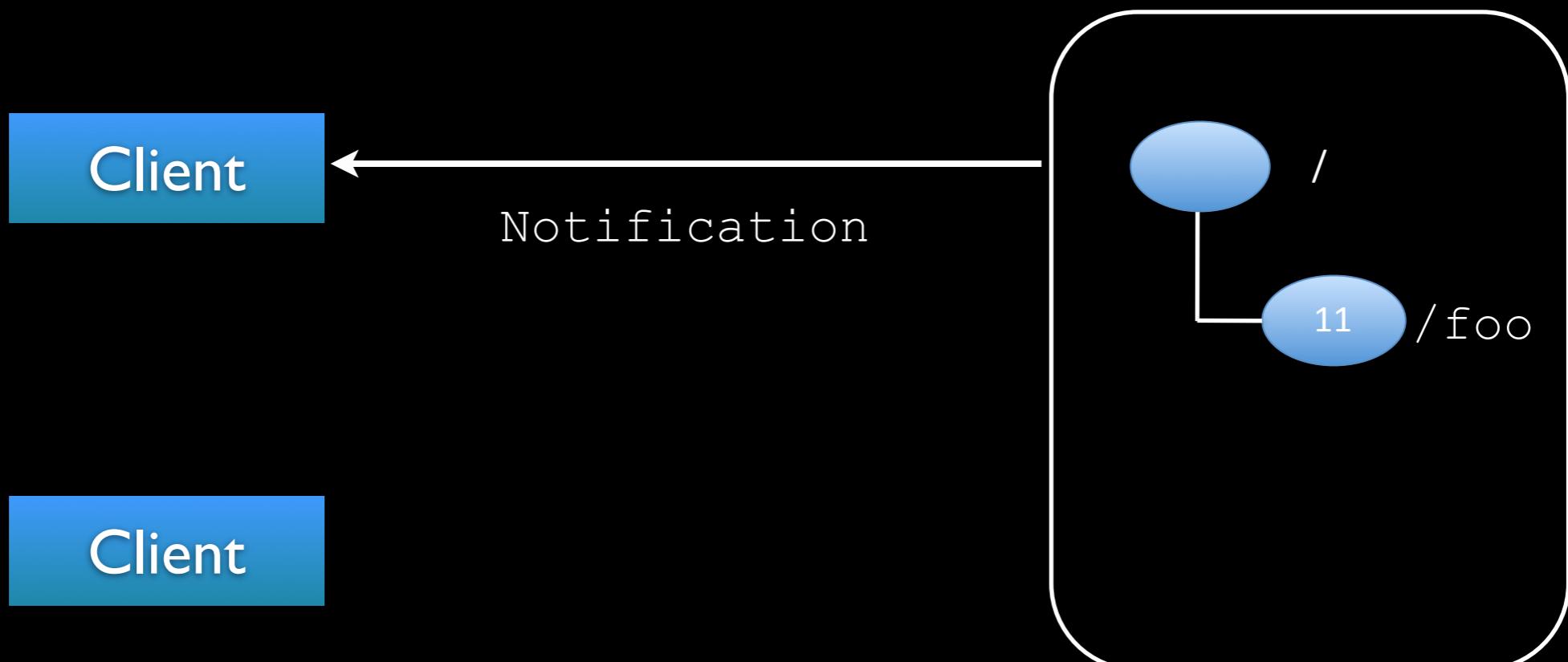
Watches



Watches



Watches

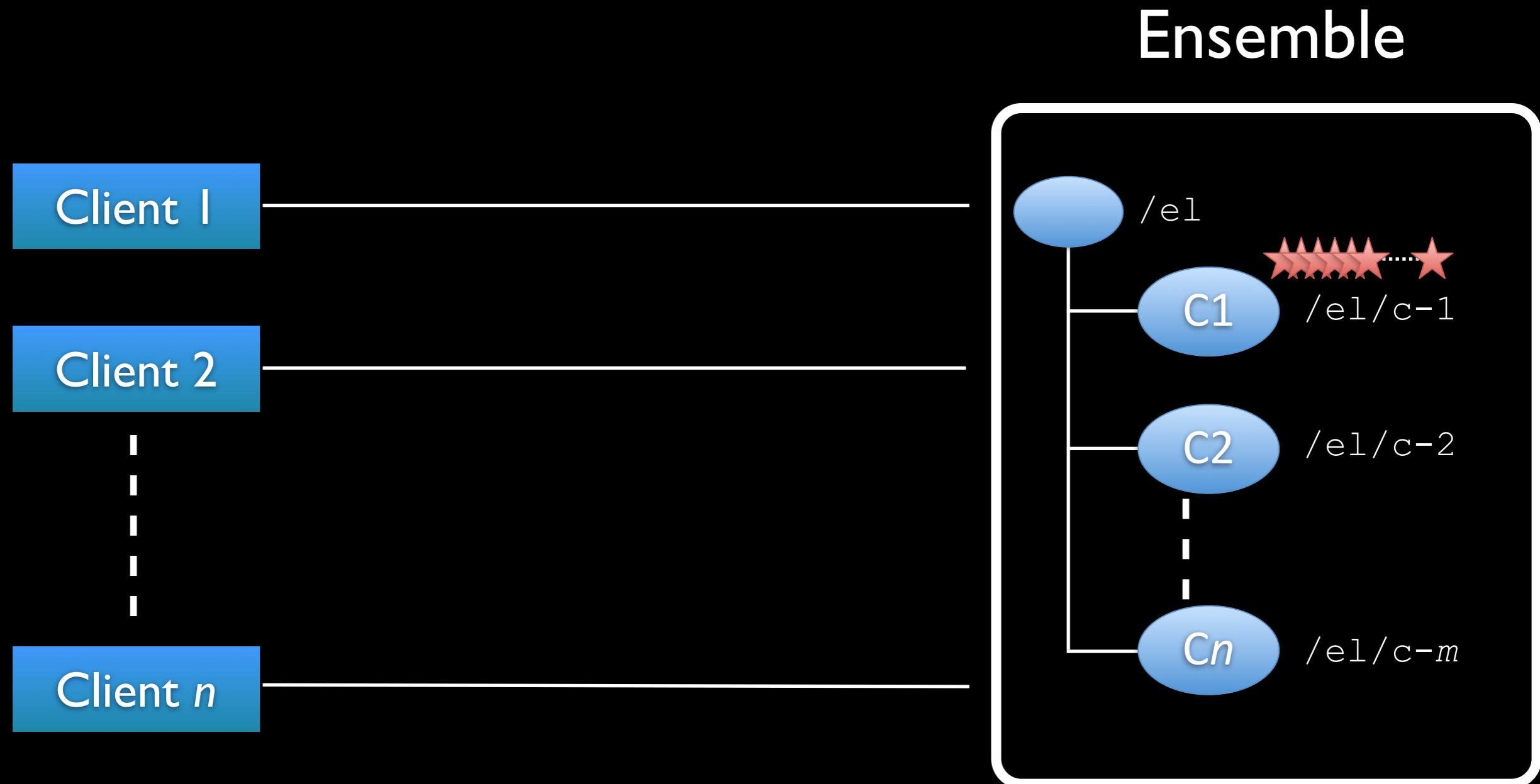


Watches, Locks, and the herd effect

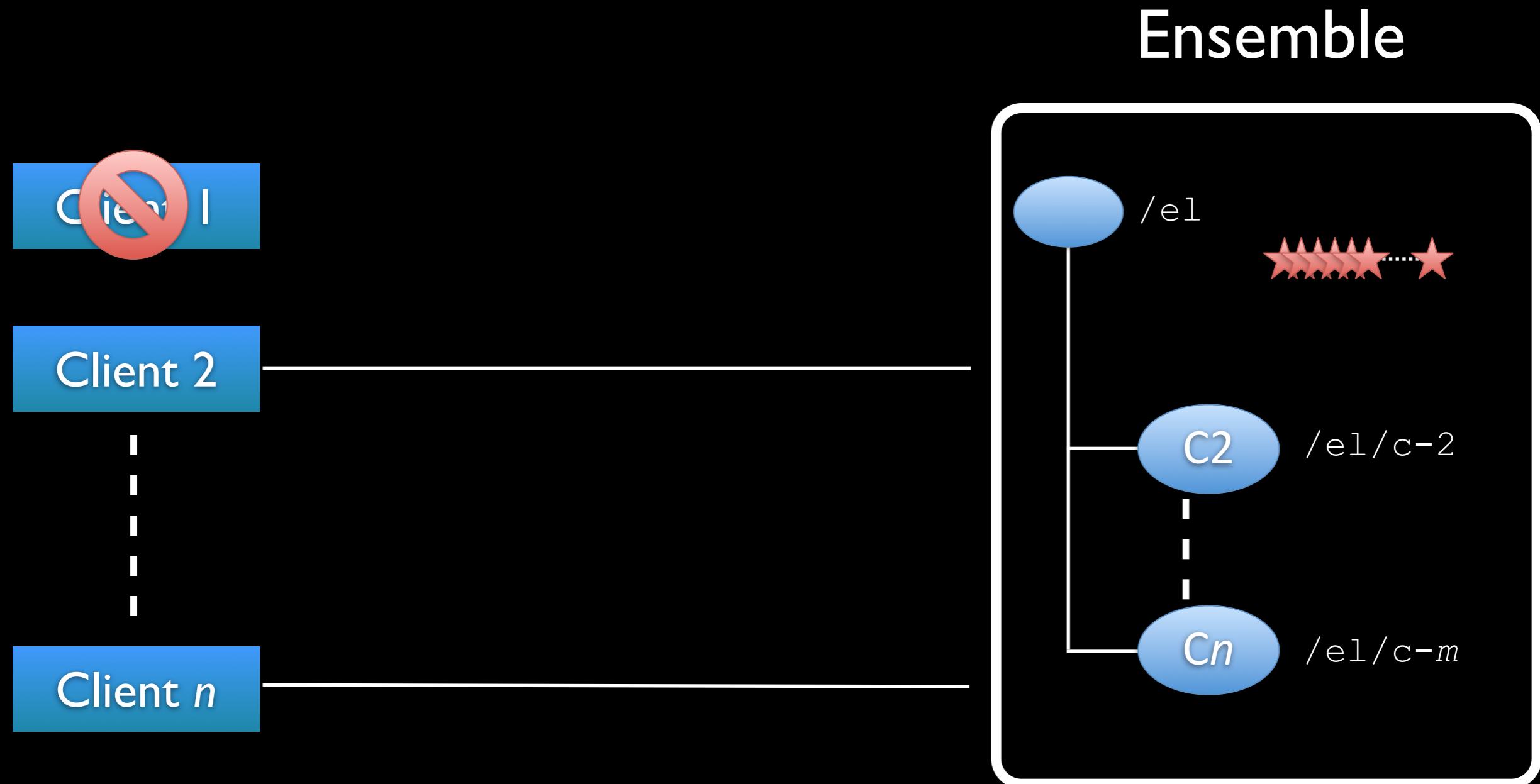
- Herd effect
 - ✓ Large number of clients wake up simultaneously
- Undesirable effect
 - ✓ Load spikes



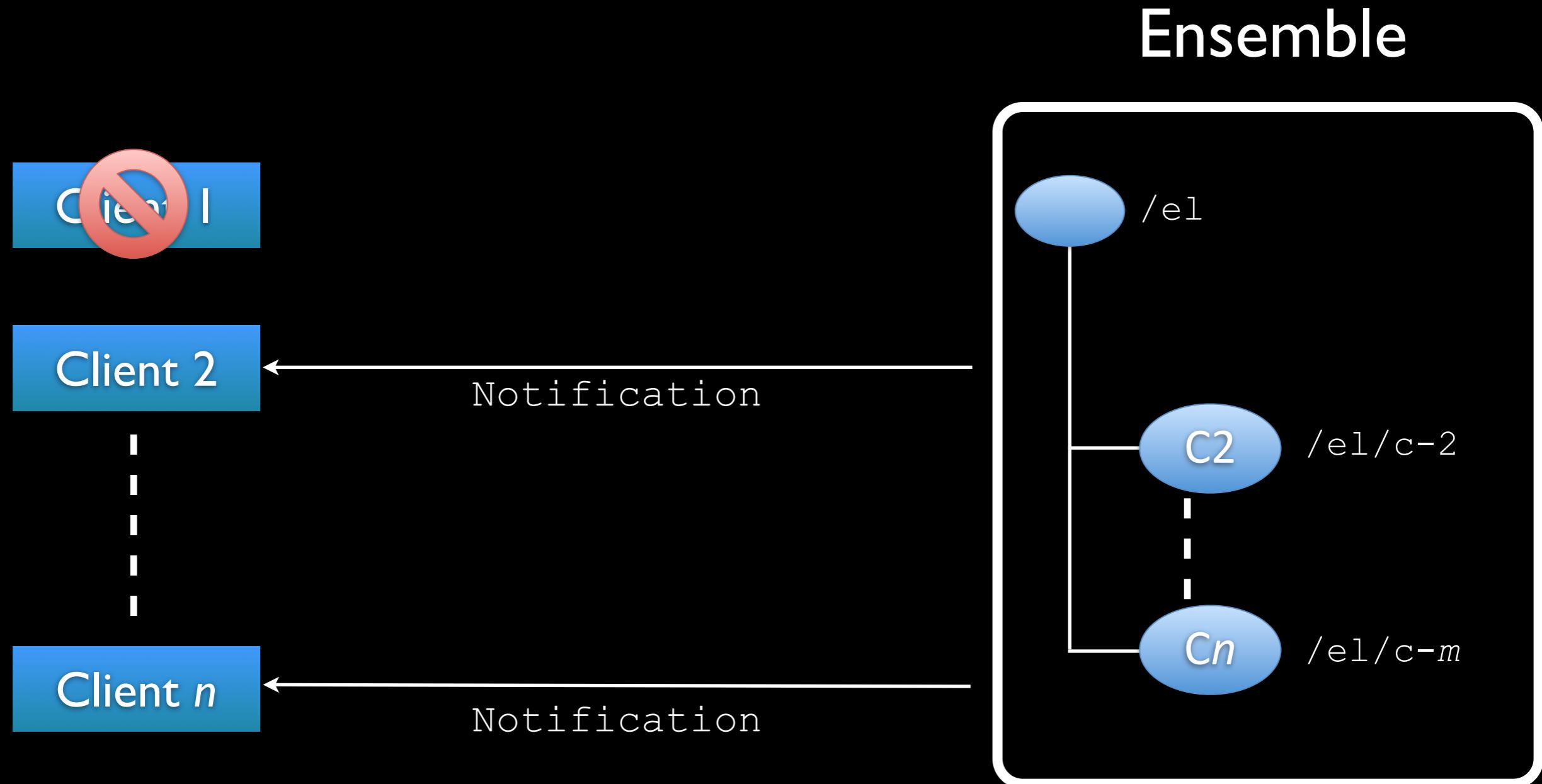
Watches, Locks, and the herd effect



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





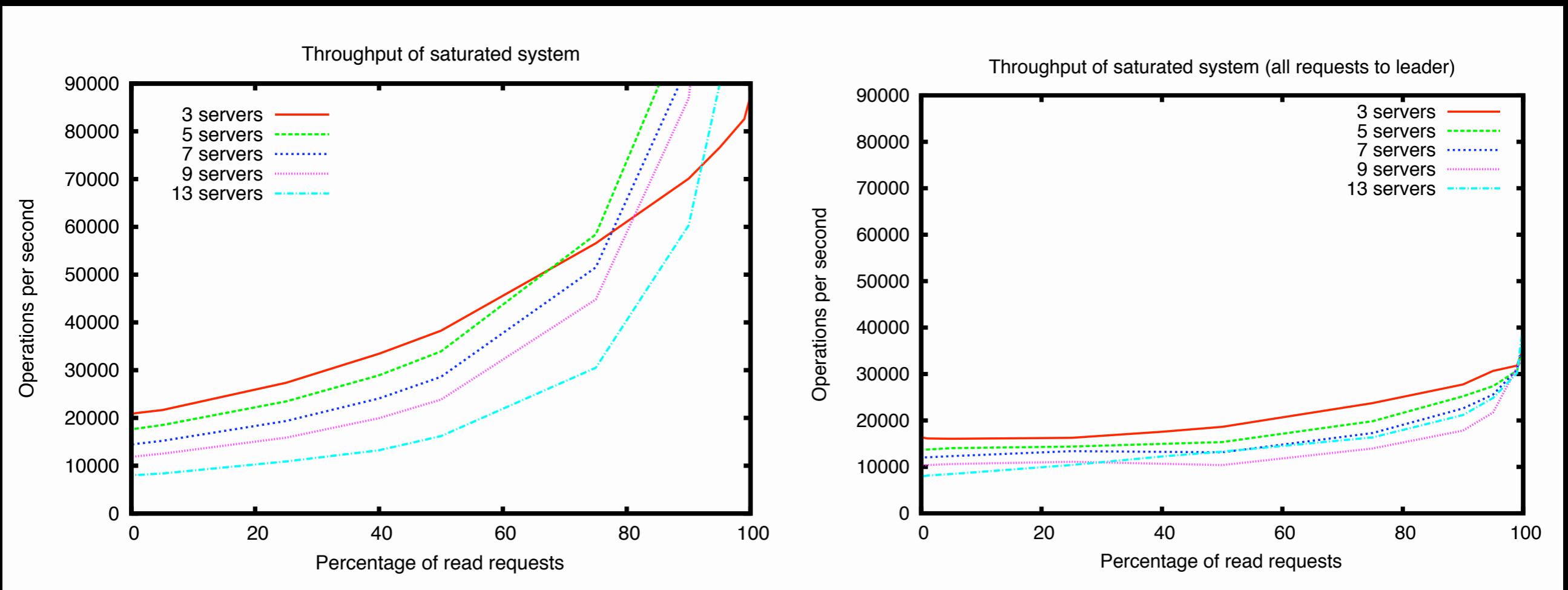
Evaluation

Evaluation

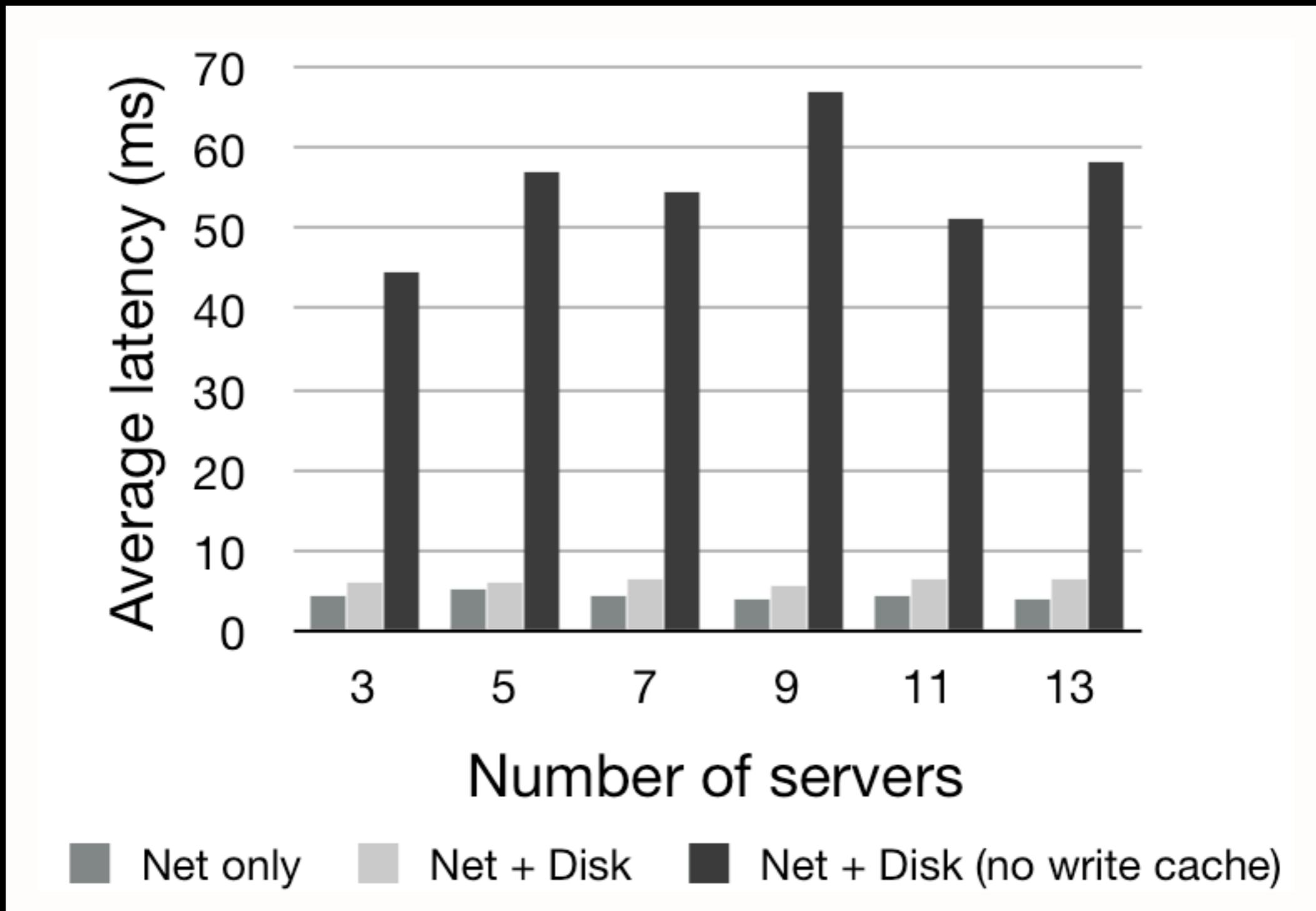
- Cluster of 50 servers
- Xeon dual-core 2.1 GHz
- 4 GB of RAM
- Two SATA disks



Throughput

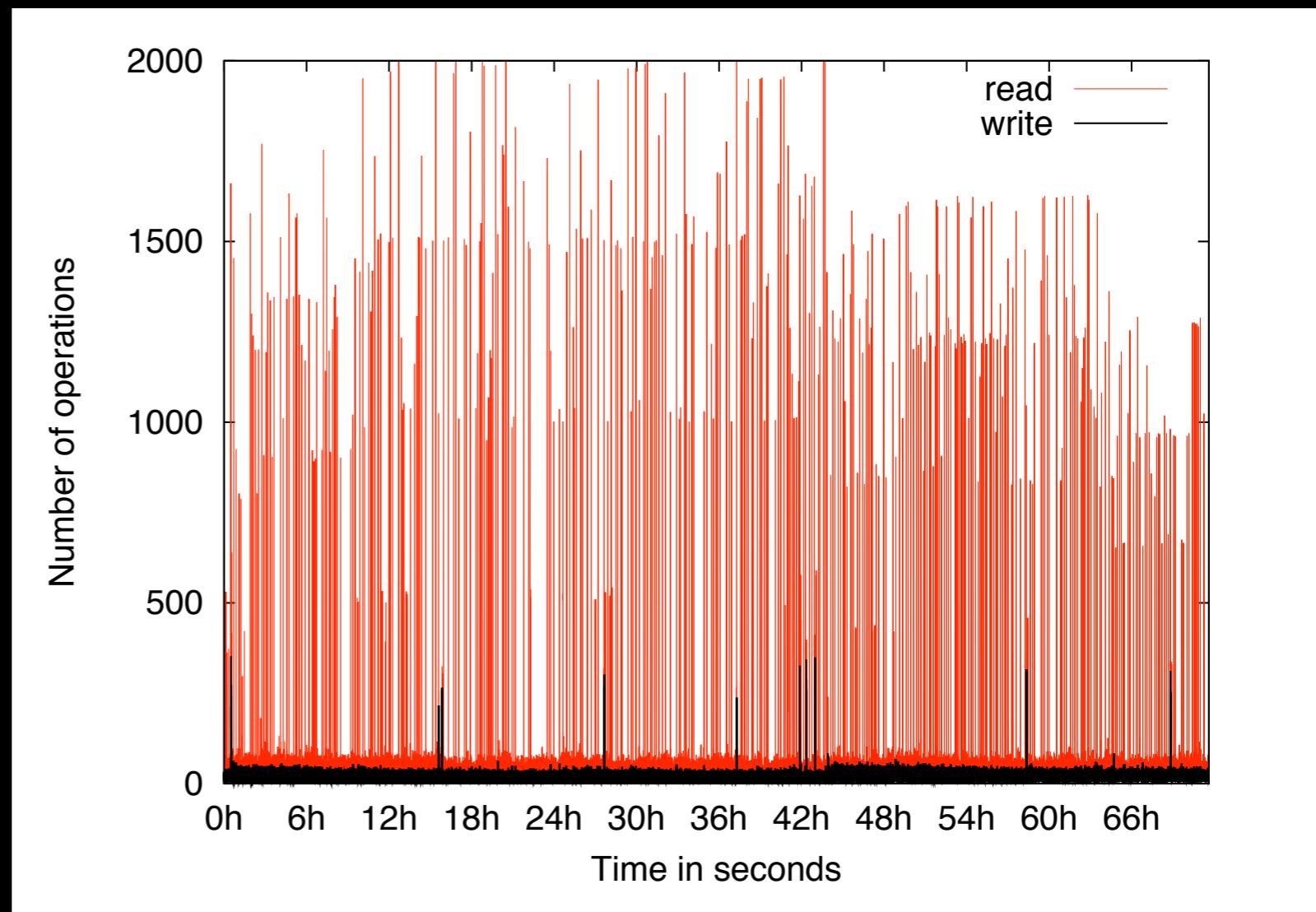


Latency



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
- Crosscolo 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>