

Jeremy Edberg

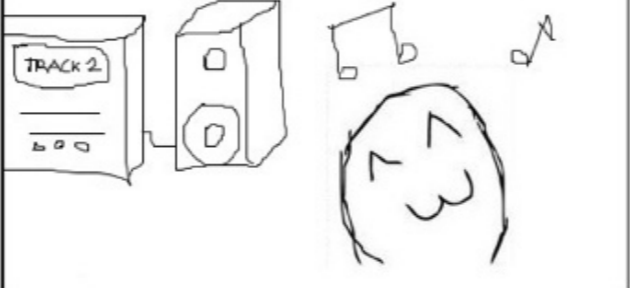
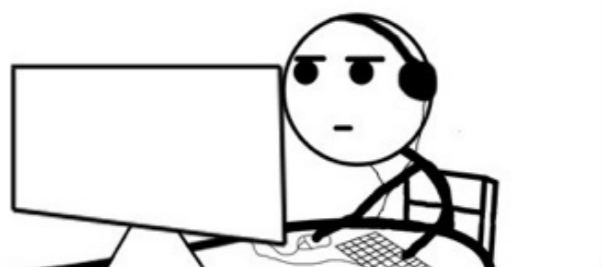
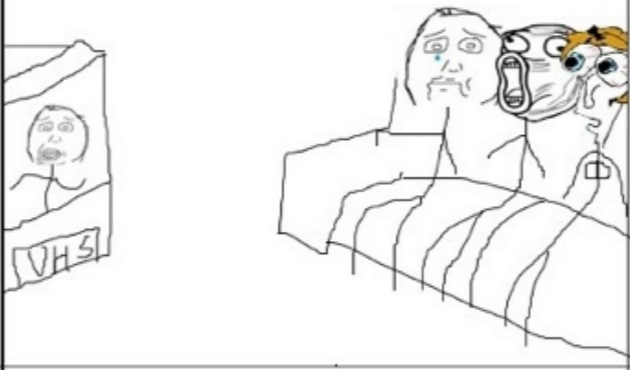
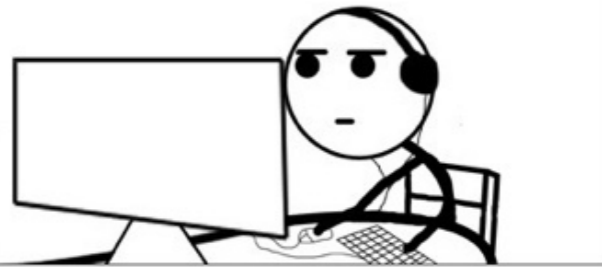

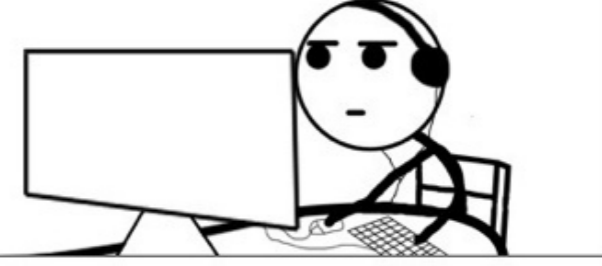



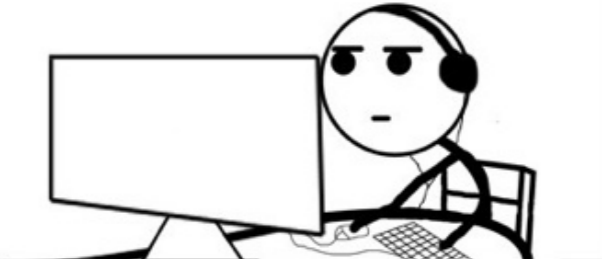
QconSF 2012



NETFLIX



NETFLIX Tweet [@jedberg](#) with feedback!

	15 Years ago	Today
Listening to music		
Watching a movie		
Contacting people		
Reading the news		
Making Music		



Building a Reliable Data Store



NETFLIX Tweet @jedberg with feedback!

Agenda

- CAP theory and how it applies to reliability
- How reddit and Netflix maintain reliable data stores
- Best Practices
- War stories -- surviving real outages



CAP Theorem

- Consistent
- Available
- Partition-resistant



ATM

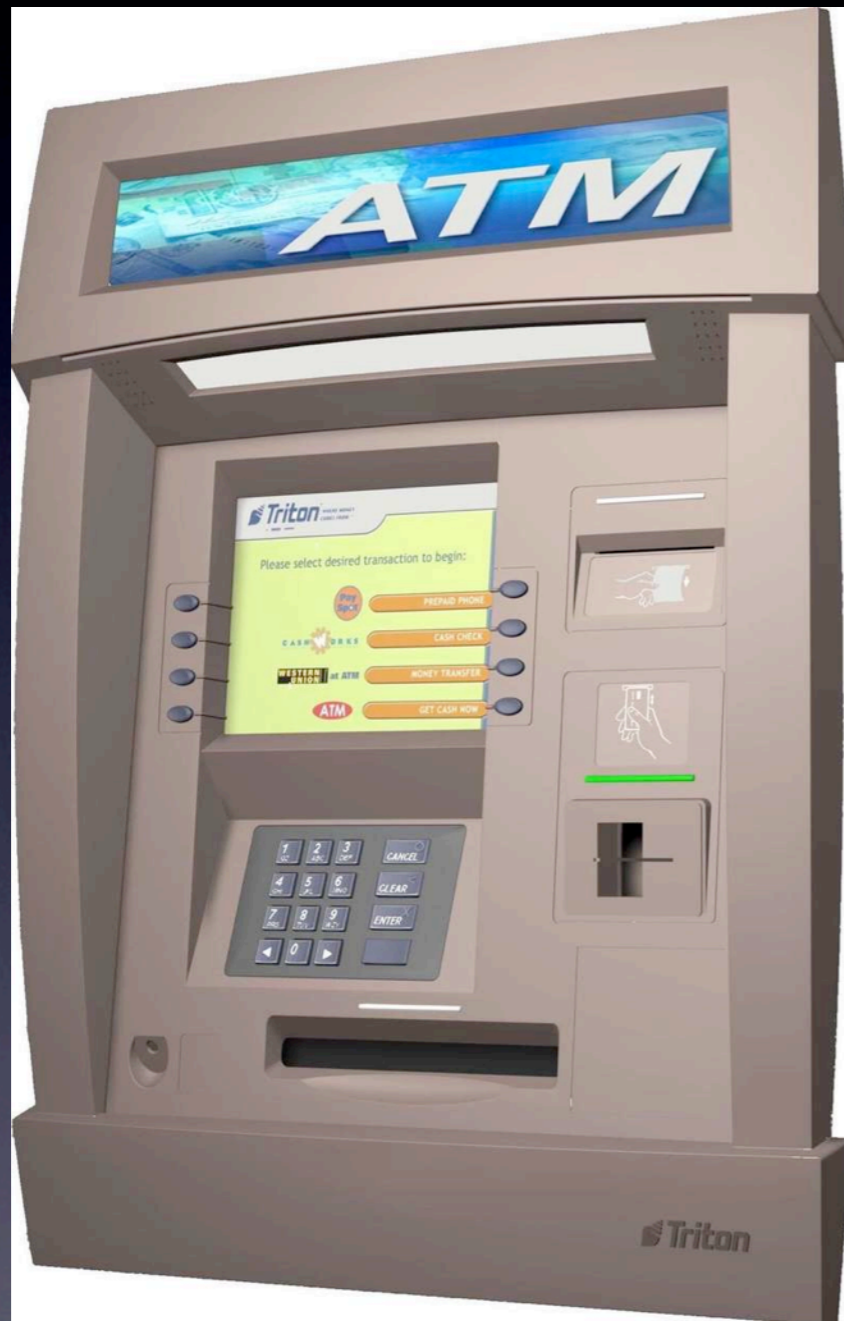


?



NETFLIX Tweet @jedberg with feedback!

ATM



AP

Limits liability through
allowing only small
transactions



NETFLIX Tweet @jedberg with feedback!

Flight Reservations



?



NETFLIX Tweet @jedberg with feedback!

Flight Reservations



AP

This is why
overbooking
occurs



NETFLIX Tweet @jedberg with feedback!



NETFLIX Tweet @jedt...

The problem with CAP

- Daniel Abadi had a problem with CAP
- The weightings were uneven
- A is essential in all scenarios
- C is more important than P
- Latency wasn't accounted for at all



PACELC

*If there is a partition (**P**) how does the system tradeoff between availability and consistency (**A** and **C**); else (**E**) when the system is running as normal in the absence of partitions, how does the system tradeoff between latency (**L**) and consistency (**C**)?*



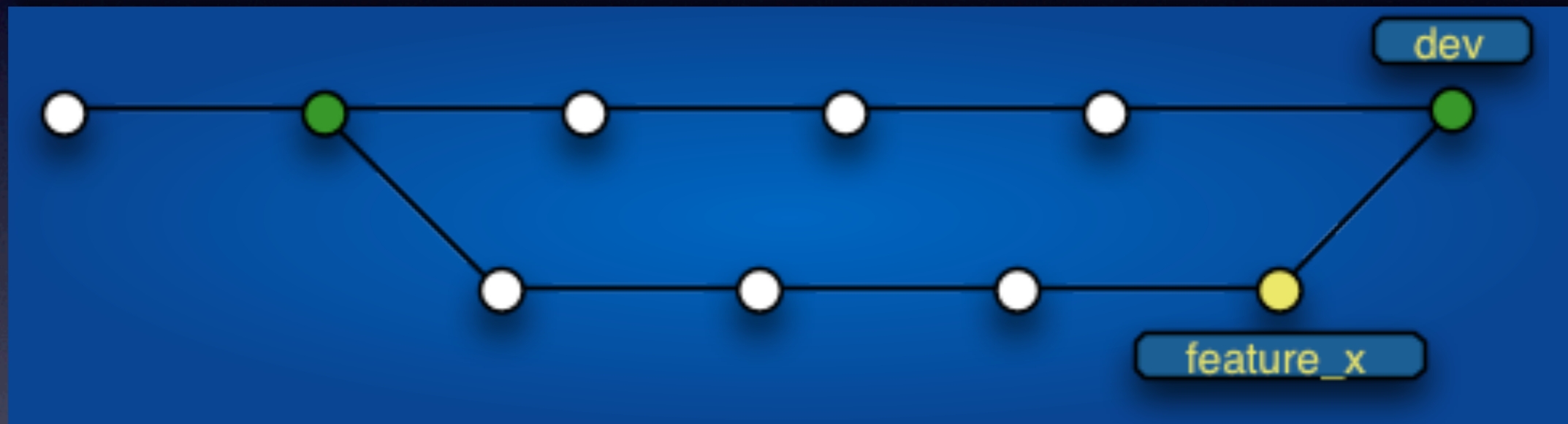
Partitioning



NETFLIX Tweet @jedberg with feedback!

Thinking like a coder

Partitions are like code branches



Some examples

- ACID systems (Postgres, Oracle, MySQL, etc) are PC/EC
- Cassandra is PA/EL



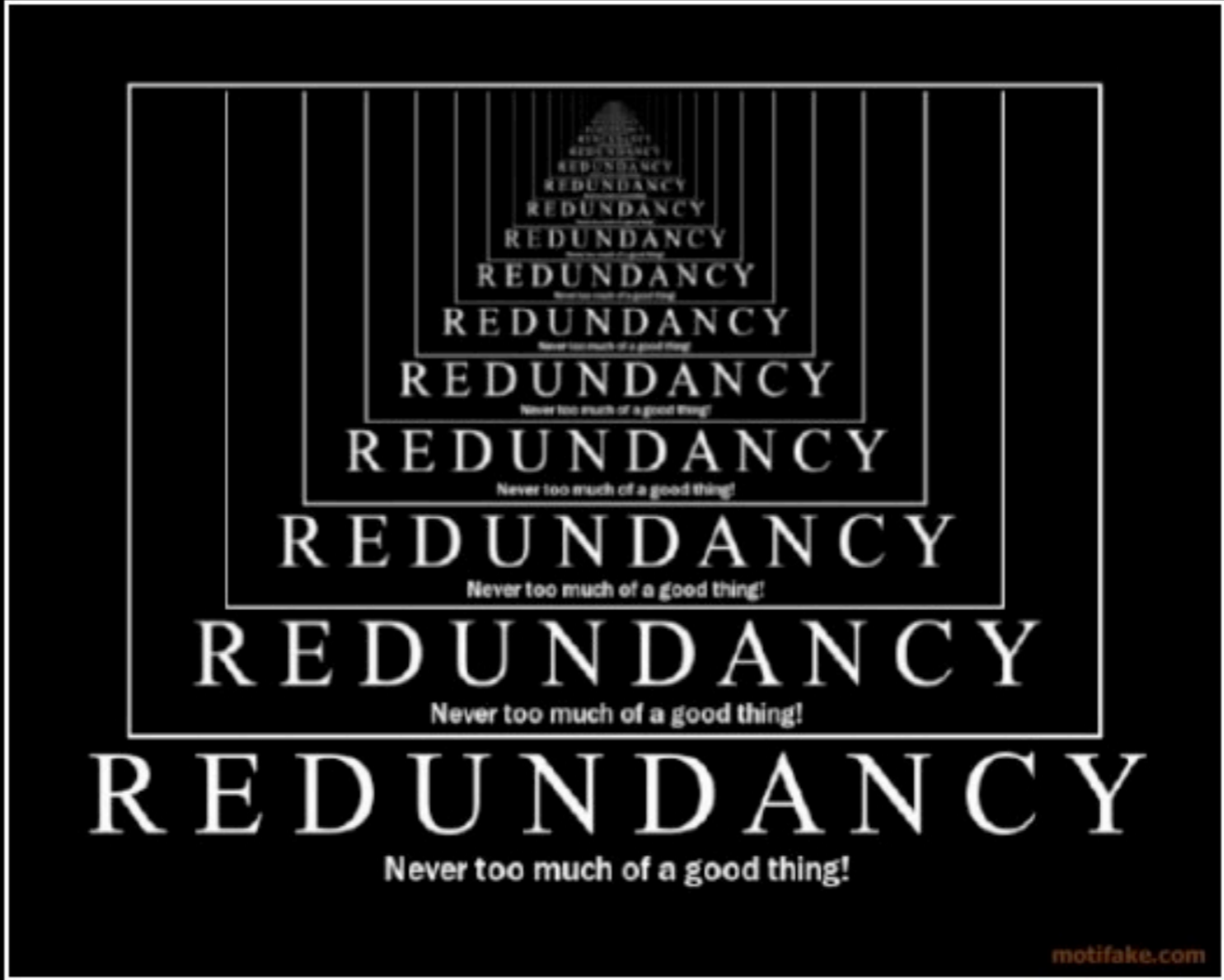


NETFLIX Tweet @jedberg with feedback!

Reliability and \$\$



Building for redundancy



Redundant Redundancy Is Redundant

motifake.com



Tweet @jedberg with feedback!

We want to make sure
we are building for
survival



NETFLIX Tweet @jedberg with feedback!

1 > 2 > 3

Going from two to three is hard



NETFLIX Tweet @jedberg with feedback!

1 > 2 > 3

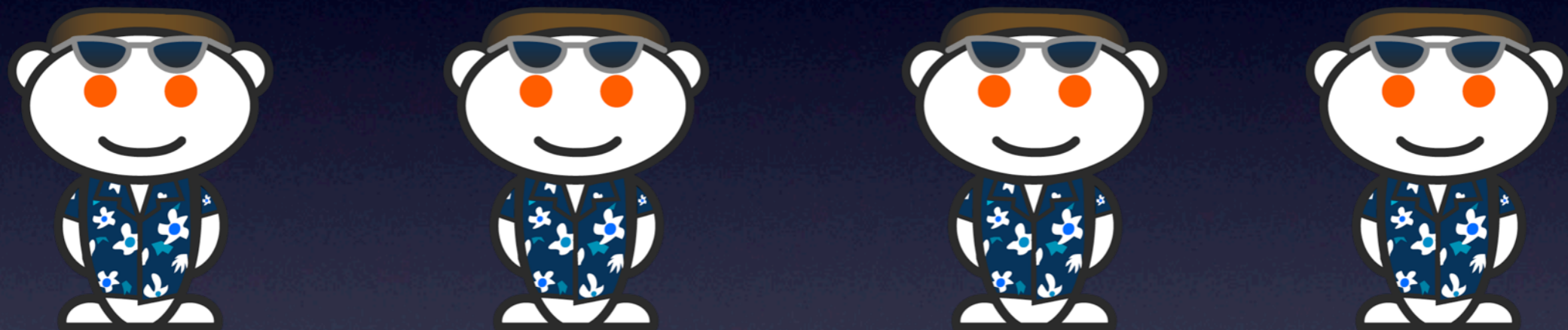
Going from one to two is harder



NETFLIX Tweet @jedberg with feedback!

Build for Three

If possible, plan for 3 or more from the beginning.



“Build for three” is the
secret to success



NETFLIX Tweet @jedberg with feedback!



NETFLIX

Tweet @jedberg with feedback!

reddit

REDDIT.COM - PICS - POLITICS - FUNNY - WTF - ASKREDDIT - SCIENCE - IAMA - PROGRAMMING - WORLDNEWS - ATHEISM - GAMING - MARIJUANA - TECHNOLOGY - DOESANYBODYELSE - COMICS - ECONOMICS - VIDEOS - OFFBEAT - EDIT »

reddit **what's hot** new controversial top saved reddit (1) | preferences | logout

reddit interviews the boy who harnessed the wind, William Kamkwamba (blog.reddit.com) promoted 1 day ago by redditads 48 comments share save hide report sponsored link what's this?

10/GUI is one of the most dramatic reimaginings of the desktop user interface I've seen in a long time. (ignorethecode.net) submitted 8 hours ago by earthboundkid to programming 584 comments share save hide report

I'm the Imgur guy, AMA! (self.IAmA) submitted 3 hours ago by MrGrim to IAmA 407 comments share save hide report

Hi Reddit!
By request and with the release of the API, I decided to do an AMA. The title says it all, so ask me anything! It doesn't have to be about imgur.
EDIT: I'll be in class for the next hour. The answers may be a little slow, but I'll do my best.

I still think this is one of the coolest videos I've ever seen. This is going on at any given moment in every living organism on the planet. (youtube.com) submitted 13 hours ago by mepardo to science 187 comments share save hide report

DNA Replication
★★★★★

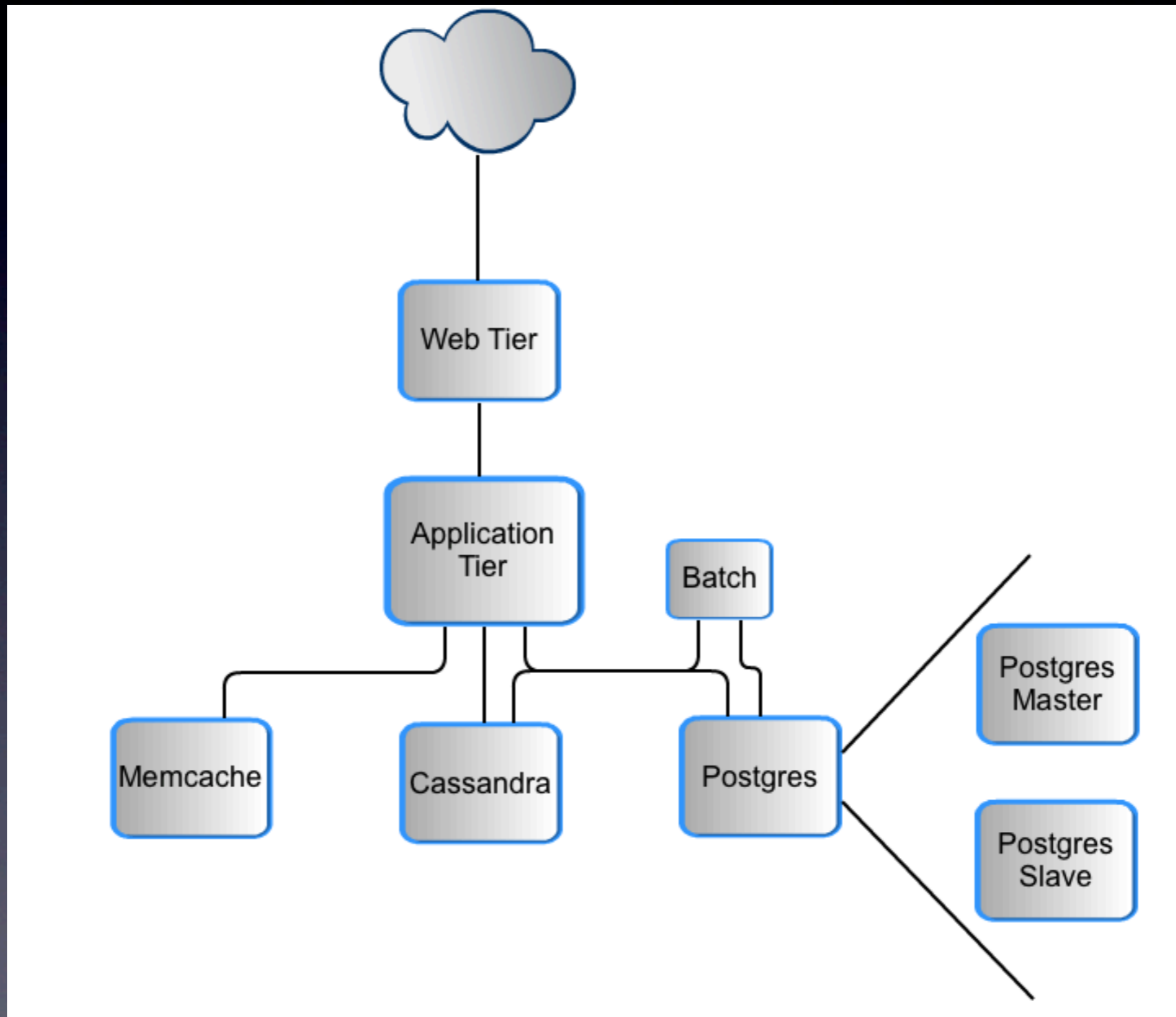
watch the reddit interview **Mike Rowe** wants to know your karma score reddit this ad

Jon Stewart spends 10 minutes ripping apart CNN's pitiful reporting [Video] (thedailyshow.com) submitted 10 hours ago by jmone to politics 205 comments share save hide report

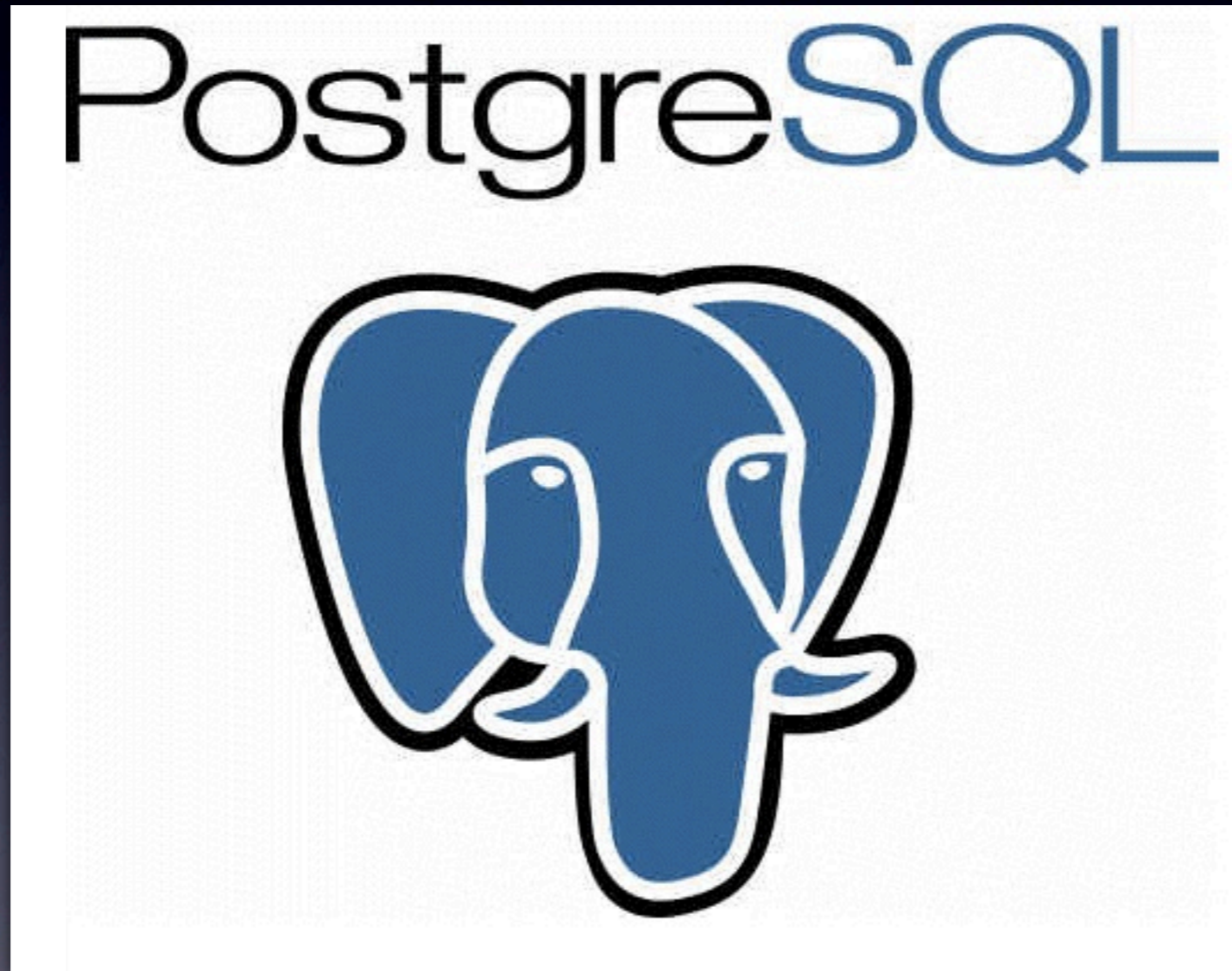


Tweet @jedberg with feedback!

Architecture



Postgres



Database Resiliency with Sharding



Sharding

- reddit split writes across four master databases
- Links/Accounts/Subreddits, Comments, Votes and Misc
- Each has at least one slave in another zone
- Avoid reading from the master if possible
- Wrote their own database access layer, called the “thing” layer



Sample Schema

```
link_thing
  int id
  timestamp date
  int ups
  int downs
  bool deleted
  bool spam
```

```
link_data
  int thing_id
  string name
  string value
  char kind
```



The thing layer

- Postgres is used like a key/value store
- Thing table has denormalized data
- Data table has arbitrary keys
- Lots of indexes tuned for our specific queries
- Thing and data tables are on the same box, but don't have to be

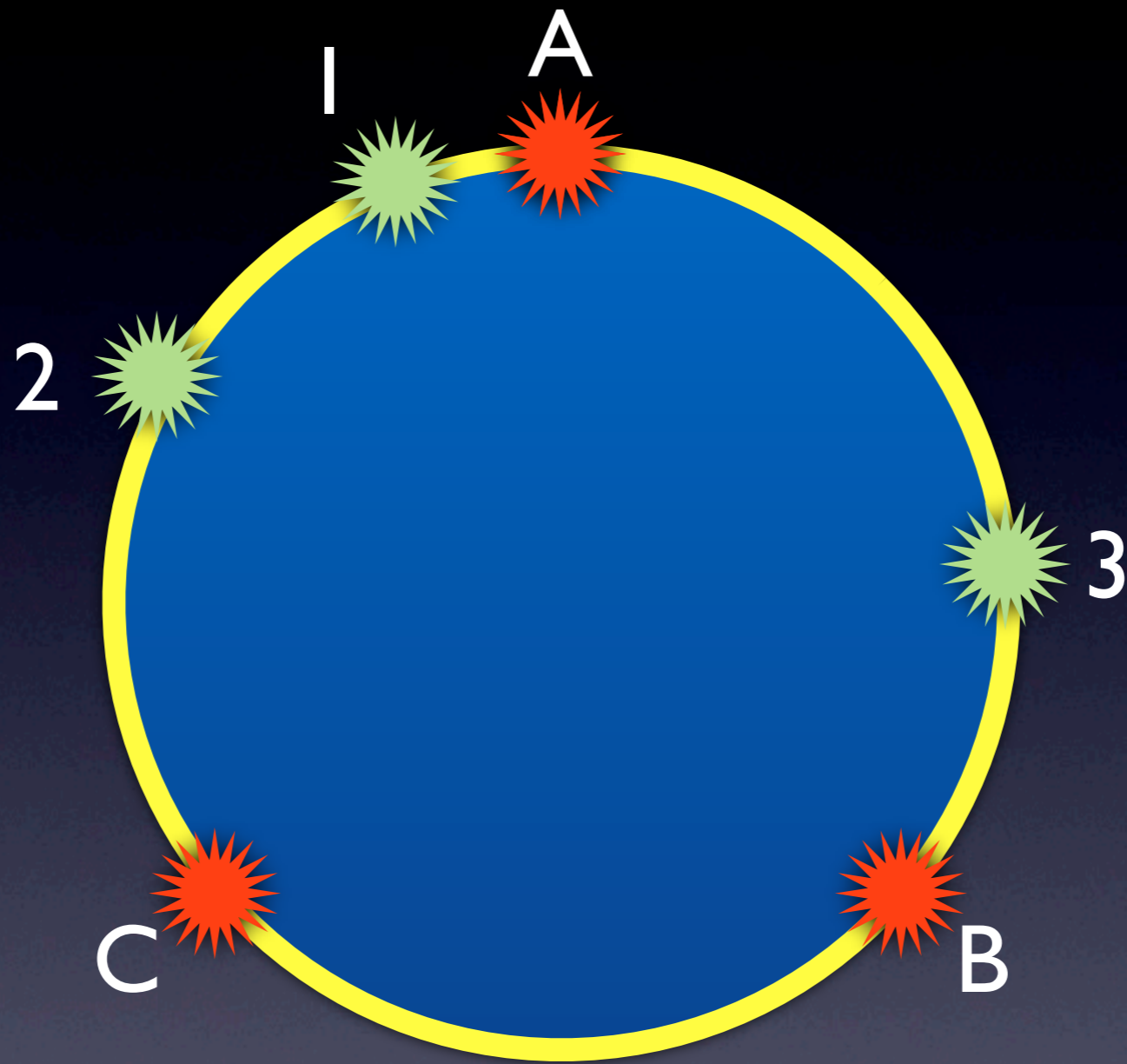


I love memcache

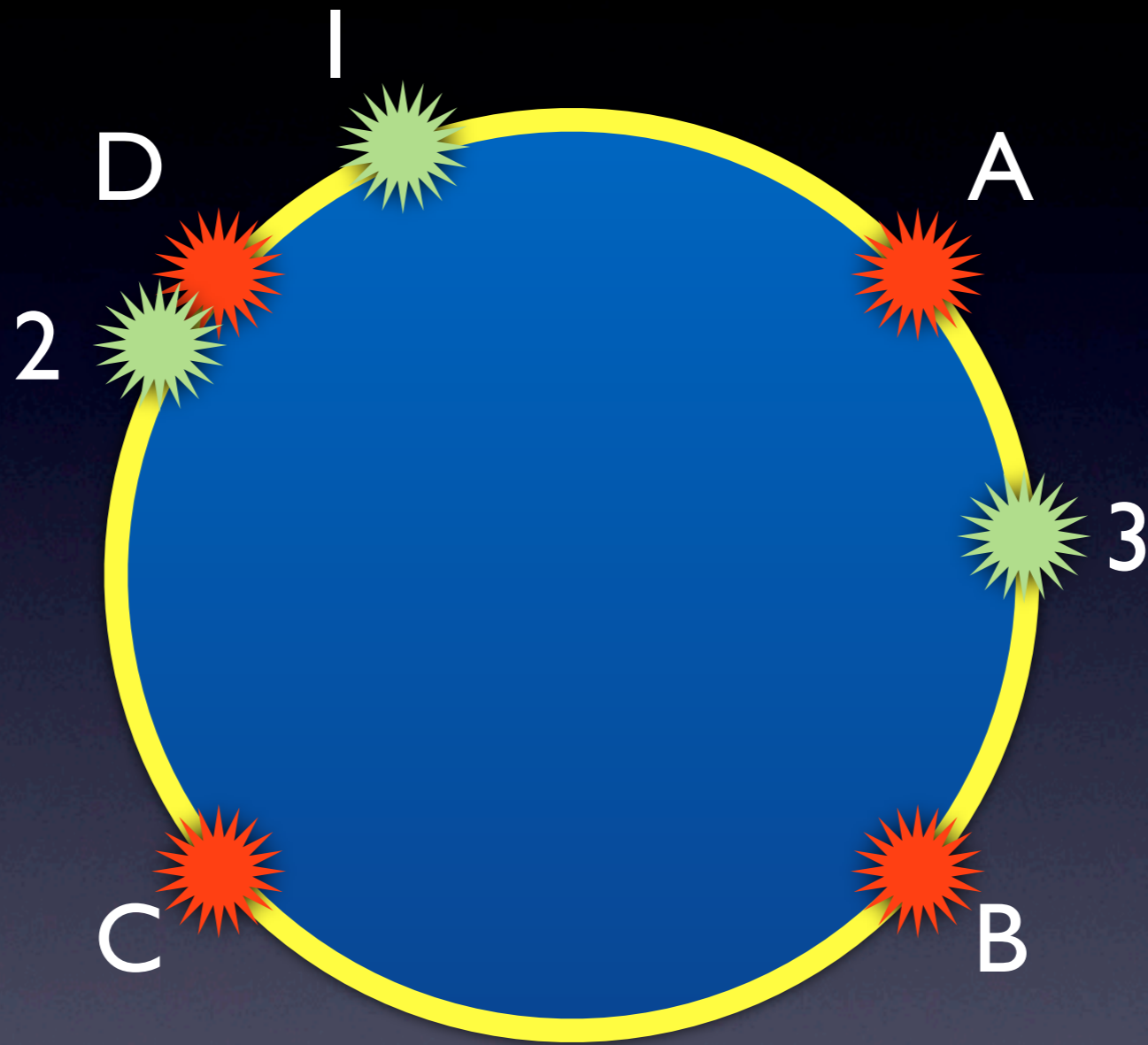
I make heavy use of memcached



NETFLIX Tweet @jedberg with feedback!

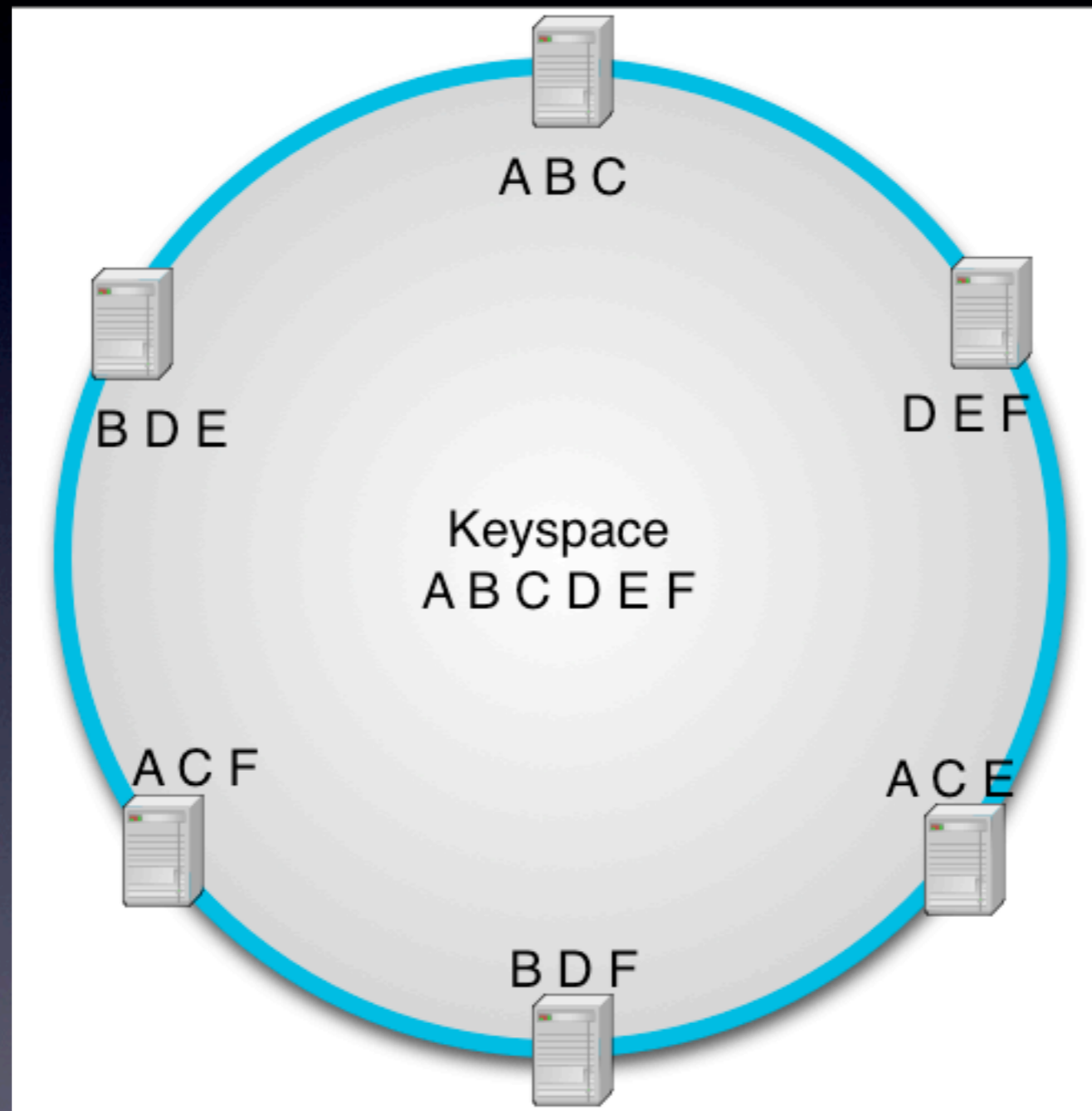


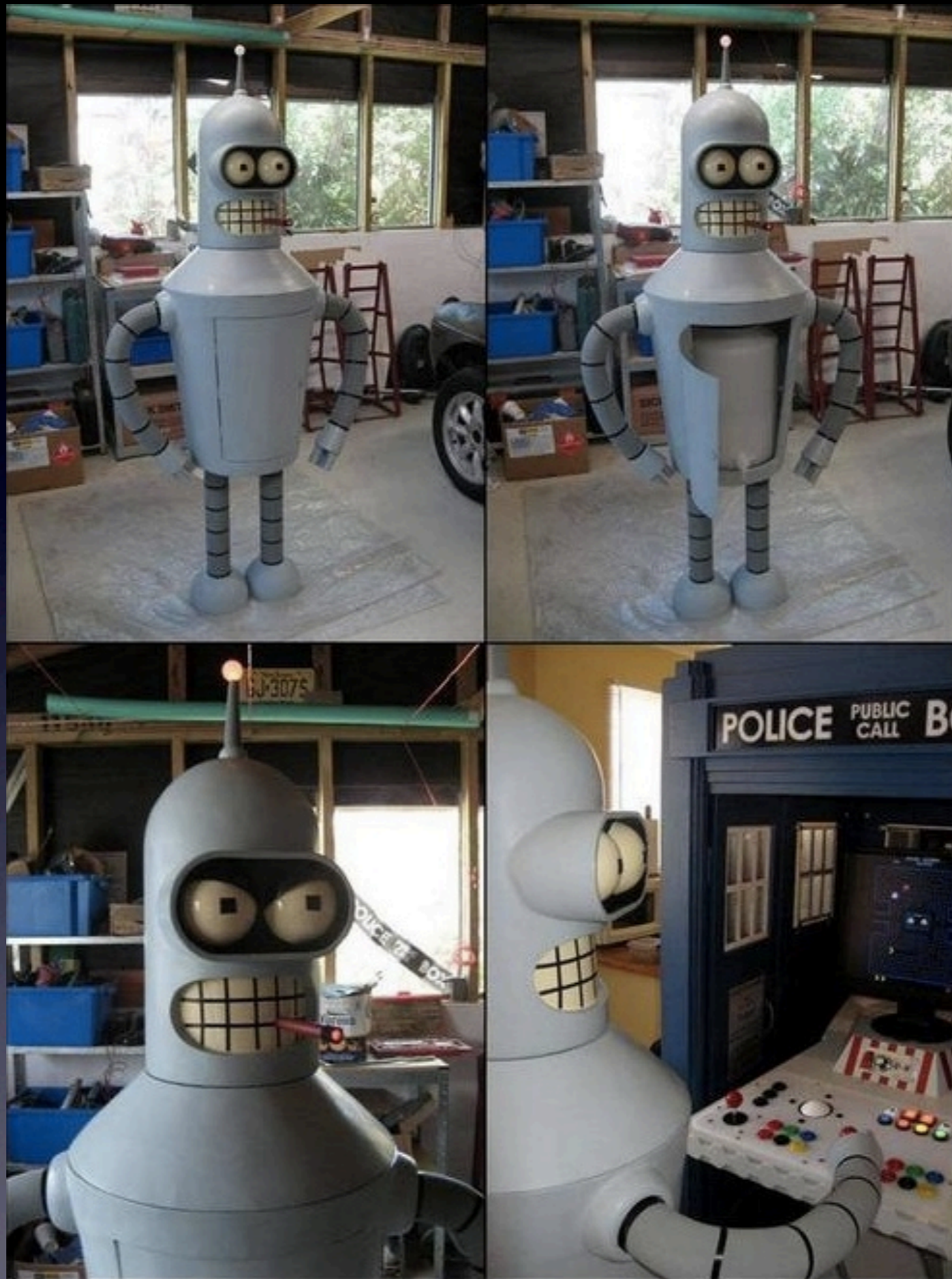
NETFLIX Tweet @jedberg with feedback!



NETFLIX Tweet @jedberg with feedback!

Cassandra





NETFLIX Tweet @jedberg with feedback!

Netflix

The screenshot shows the Netflix homepage with a red header containing the logo and user information. Below the header are navigation tabs for 'Watch Instantly', 'Just for Kids', 'Browse DVDs', 'Your Queue', and 'Suggestions For You'. A search bar is located on the right. The main content is organized into three sections: 'New movies to watch instantly' (featuring titles like 'Expendables', 'Dazed and Confused', 'OMG!', 'The Top 50 Incidents in WWE History', 'H30', 'Cypher', 'Love Wedding Marriage', and 'Walk on the Moon'), 'New TV to watch instantly' (featuring 'South Park', 'The Backyardigans', 'Breaking Bad', 'Mad Men', 'Grey's Anatomy', 'Desperate Housewives', 'Pawn Stars', and 'Star Trek'), and 'Watch It Again' (featuring '30 Rock', 'Pulp Fiction', 'Apollo 13', 'The Hunt for Red October', 'Spacemans', 'Being John Malkovich', 'Futurama', and 'Hot Time in the Heart of London').



NETFLIX Tweet @jedberg with feedback!

Data

What does Netflix do with it all?



NETFLIX Tweet @jedberg with feedback!

We store it!

- Cache (memcached)
- Cassandra
- RDS (MySql)



I love memcache

I make heavy use of memcached



NETFLIX Tweet @jedberg with feedback!

RDS (Relational Database Service)

The screenshot shows the Netflix website interface. At the top, the Netflix logo is on the left, and the user's name "Jeremy Edberg" and "Your Account & Help" are on the right. Below the logo, there are navigation tabs: "Watch Instantly", "Just for Kids", "Browse DVDs", "Your Queue", and "Taste Profile". A search bar is located to the right of these tabs, containing the text "Movies, TV shows, actors, directors, genres" and a search icon. Below the navigation tabs, there are links for "Genres", "New Arrivals", and "Instantly to your TV".

The main content area is divided into two sections: "Recently Watched" and "Top 10 for Jeremy".

Recently Watched: The first item is the movie "IRON MAN 2".

Top 10 for Jeremy: The first item is the TV show "Mad Men". A tooltip is displayed over this item, providing the following information:

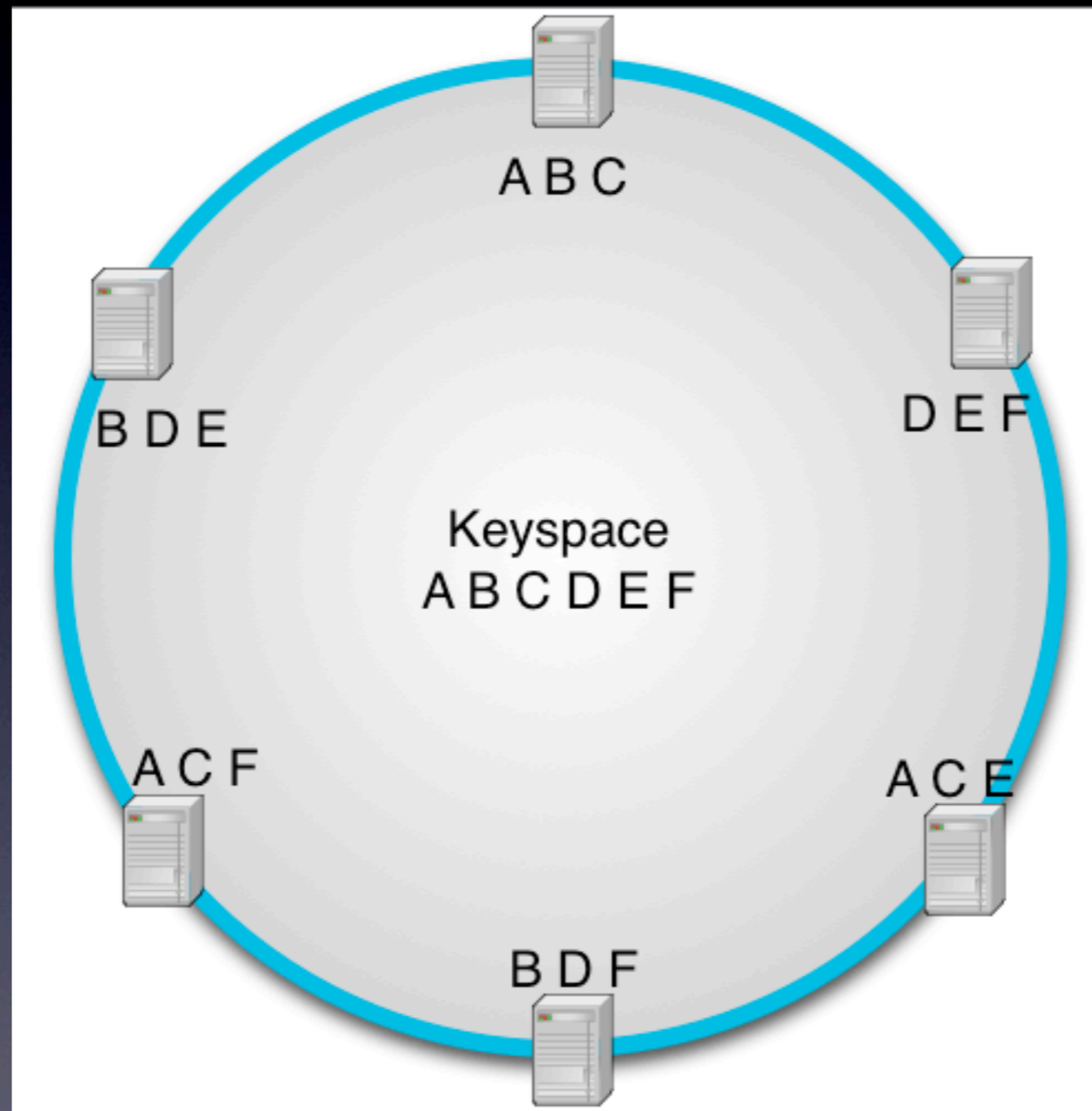
- Mad Men**
- 2007-2010 TV-14 Seasons 1-4
- Set in 1960s New York City, this series takes a peek inside an ad agency during an era when the cutthroat business had a glamorous lure. [More Info](#)
- Starring: Jon Hamm, Elisabeth Moss
- Creator: Matthew Weiner
- Our best guess for Jeremy: ★★★★★
- Buttons: "Not Interested" and "+ Instant Queue"

Other items in the "Top 10 for Jeremy" list include "Breaking Bad", "the Office", and "AMERICAN DAD!".

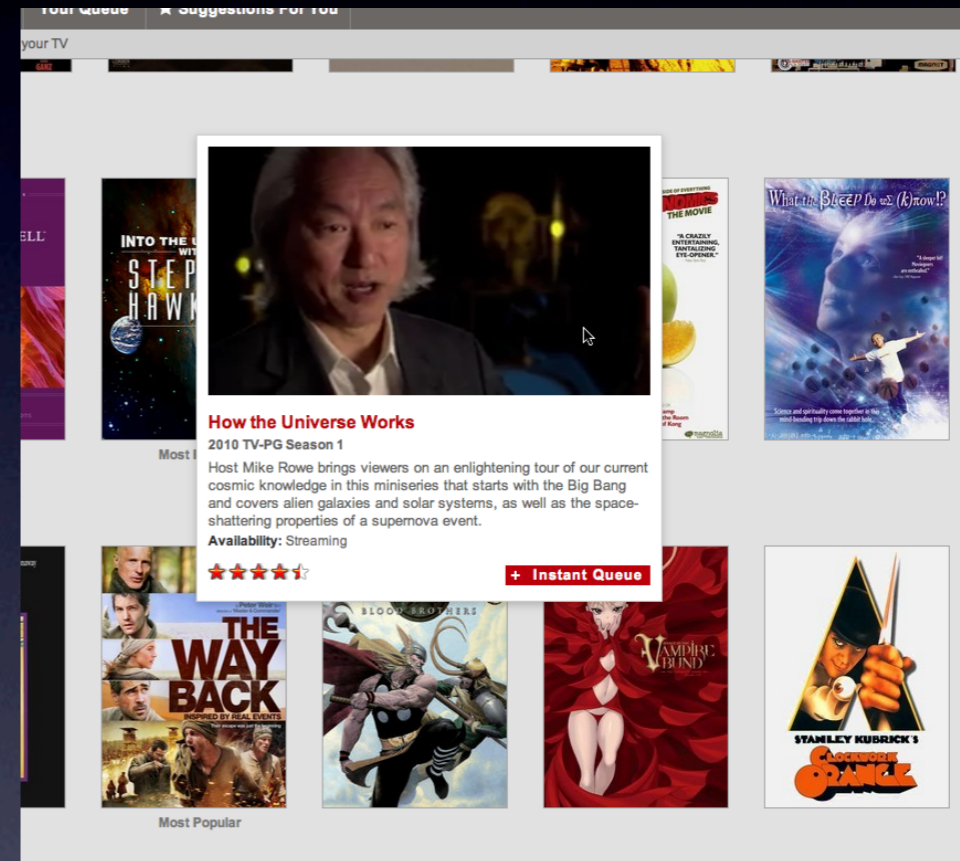
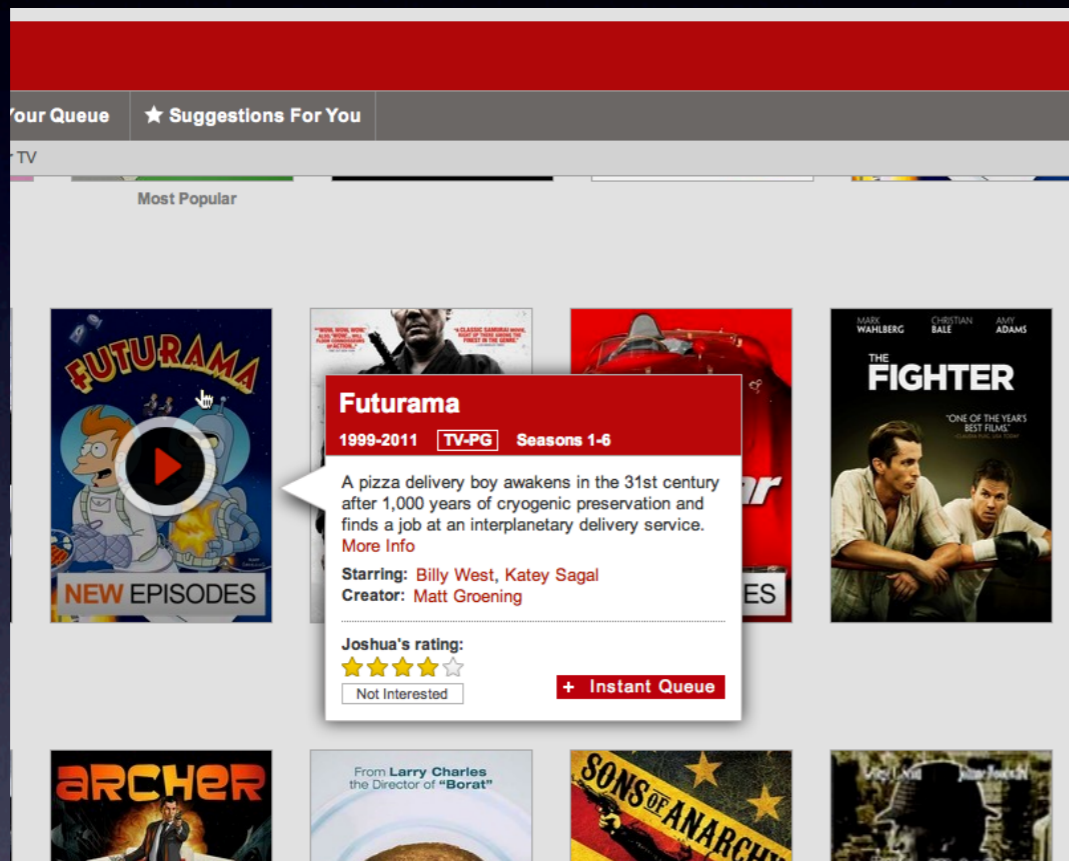


NETFLIX Tweet @jedberg with feedback!

Cassandra



A/B Testing



NETFLIX Tweet @jedberg with feedback!

A/B Testing

Online Data	Offline Data
Test Cell allocation Test Metadata Start/End date UI Directives	Test tracking Retention Fraction Viewed Pages Viewed



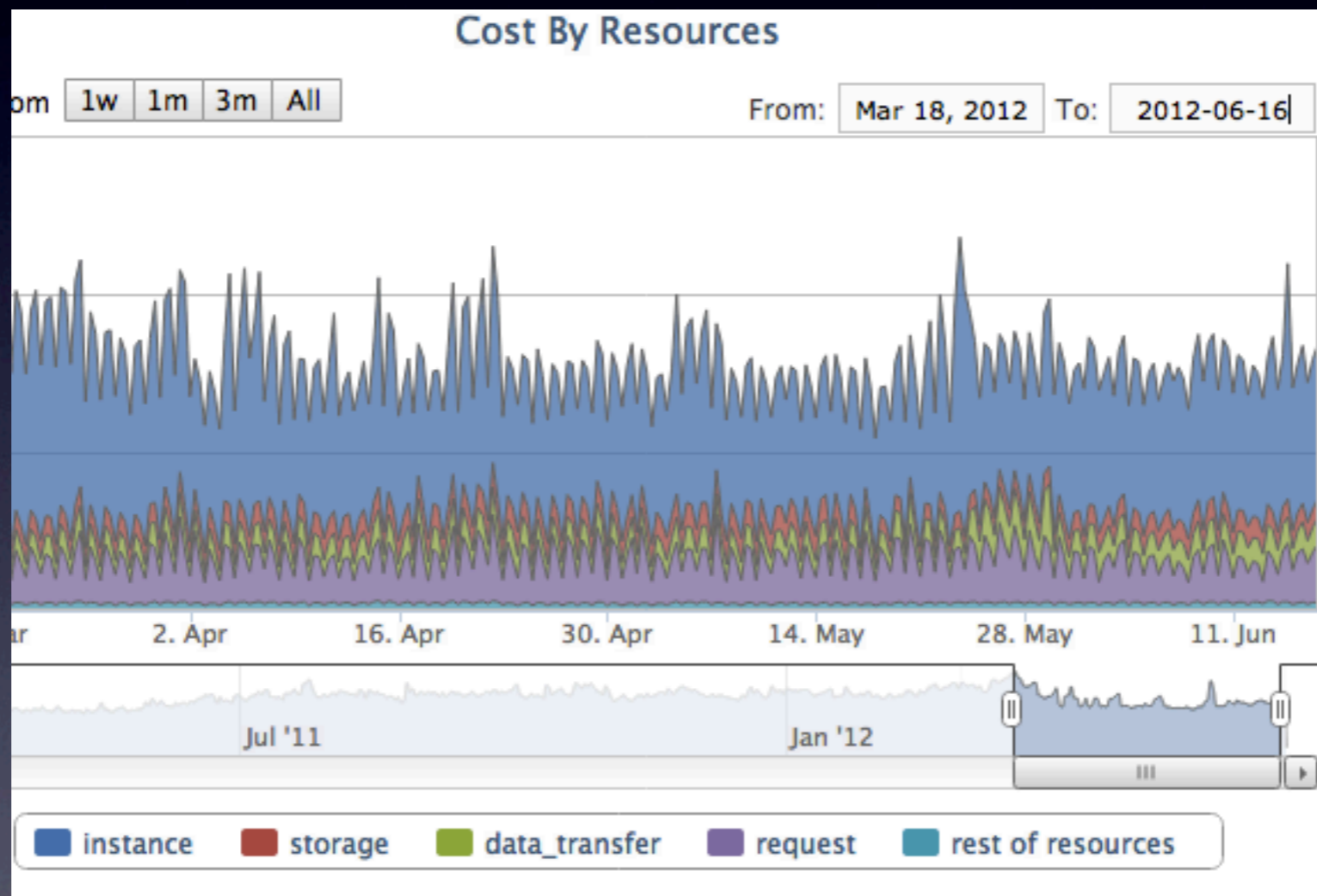
Atlas



NETFLIX Tweet @jedberg with feedback!

AWS Usage

Dollar amounts have been carefully removed



NETFLIX Tweet @jedberg with feedback!

Chronos

Search

Tag: Values:

Time Period:

Tag:

- all
- app
- batch
- cluster
- cmc
- environment
- region
- source

Results

Total Events: 92

[Permalink](#)

Timestamp (PDT)	Application	Event Type	Region	CMC	CMC Details	Description
2012-06- T11:35:38	cass_xxx	SecurityGroup	us-east-1		No CMC Details	changed SG cass_xxx
2012-06- T11:35:16	AWS	EnableInstance	us-east-1		No CMC Details	Enable _____ in Discovery
2012-06- T11:33:01	nccp	TerminateInstance	us-east-1		No CMC Details	Terminate instance _____ and shrink auto scaling group 'nccp-debug'
2012-06- T11:32:59	nccp	UpdateASG	us-east-1		No CMC Details	Update Autoscaling Group 'nccp-debug'
2012-06- T11:32:29	apidaemon	DeleteASG	us-east-1		CMC Details	Deleting apidaemon
2012-06- T11:30:37	cass_xxx	SecurityGroup	us-east-1		No CMC Details	changed SG cass_xxx
2012-06- T11:28:47	AWS	DisableInstance	us-east-1		No CMC Details	Disable _____ in Discovery
2012-06- T11:28:09	AWS	EnableInstance	us-east-1		No CMC Details	Enable _____ in Discovery
2012-06- T11:26:18	AWS	DisableInstance	us-east-1		No CMC Details	Disable _____ in Discovery
2012-06- T11:25:45	beaconserver	UpdateASG	eu-west-1		CMC Details	Update Autoscaling Group 'beaconserver-pt-v002'
2012-06- T11:25:39	AWS	EnableInstance	us-east-1		No CMC Details	Enable _____ in Discovery
2012-06- T11:25:35	cass_xxx	SecurityGroup	us-east-1		No CMC Details	changed SG cass_xxx
2012-06- T11:25:20	beaconserver	UpdateASG	eu-west-1		CMC Details	Update Autoscaling Group 'beaconserver-ce-v001'
2012-06- T11:25:02	beaconserver	UpdateASG	eu-west-1		CMC Details	Update Autoscaling Group 'beaconserver-v021'
2012-06- T11:23:07	MerchWeb	CreateFastProperty	us-east-1		CMC Details	Create Fast Property



More Things Netflix Stores in Cassandra

- Video Quality
- Network issues
- Usage History
- Playback Errors

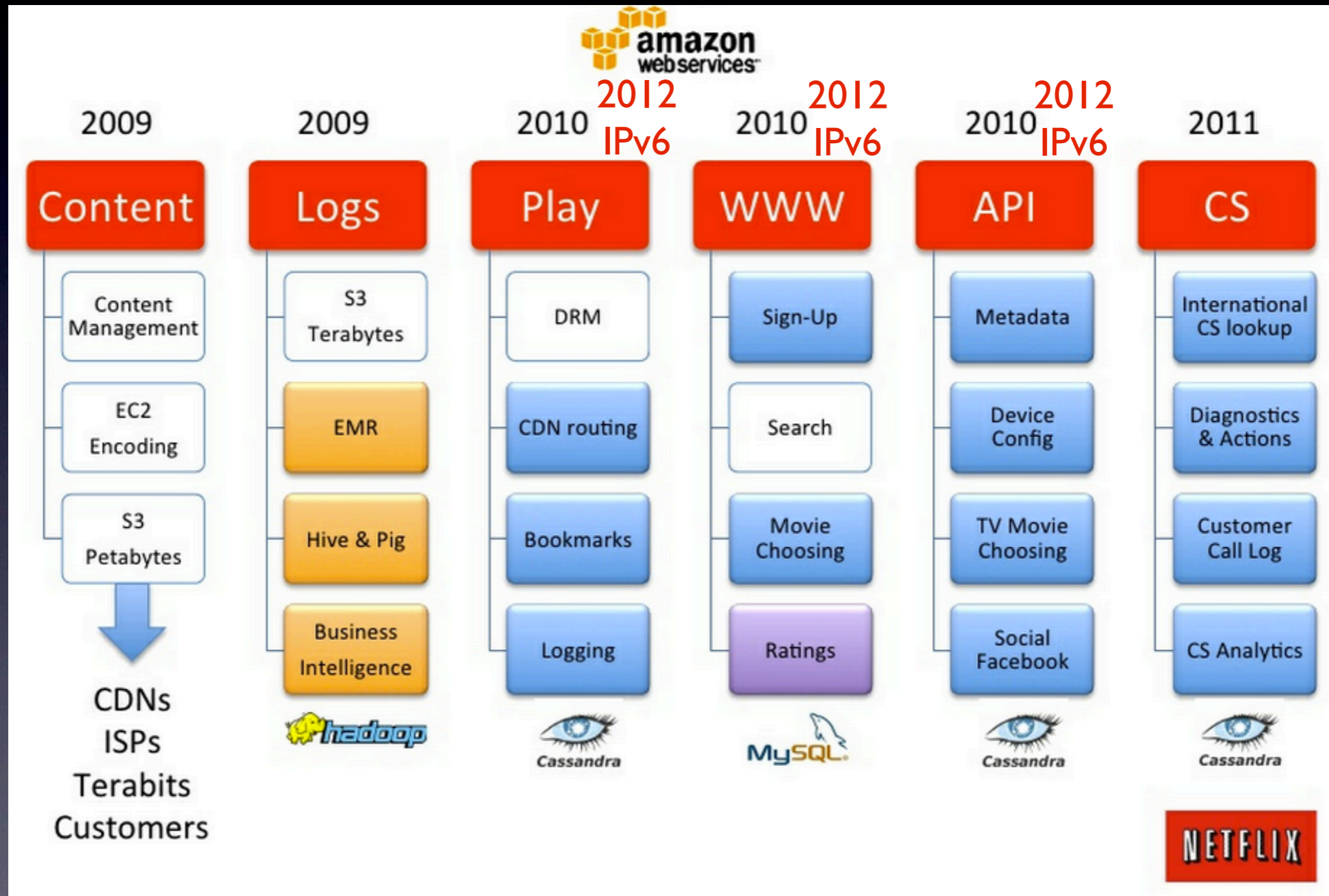


NETFLIX Tweet @jedberg with feedback!

Service based architecture



Netflix on AWS



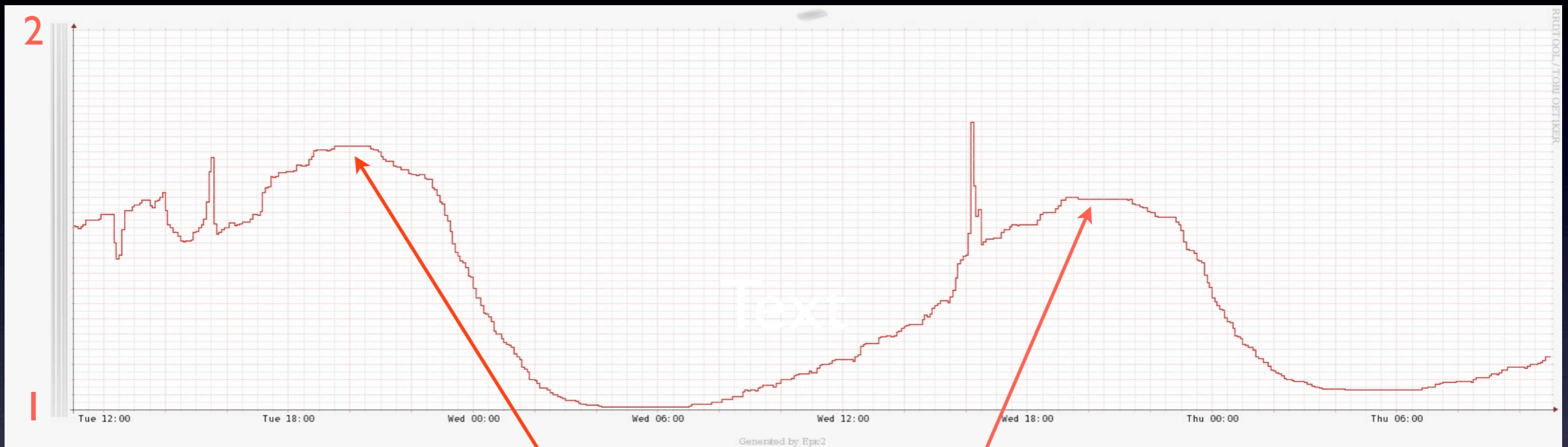
Tweet @jedberg with feedback!

Abstraction

- Data sources are abstracted away behind restful interfaces
- Each application owns its own consistency
- Each application can scale independently based on load



Netflix autoscaling



Traffic Peak



NETFLIX Tweet @jedberg with feedback!

The Big Oracle Database



NETFLIX Tweet @jedberg with feedback!

Circuit Breakers

Be liberal in what you accept, strict in what you send

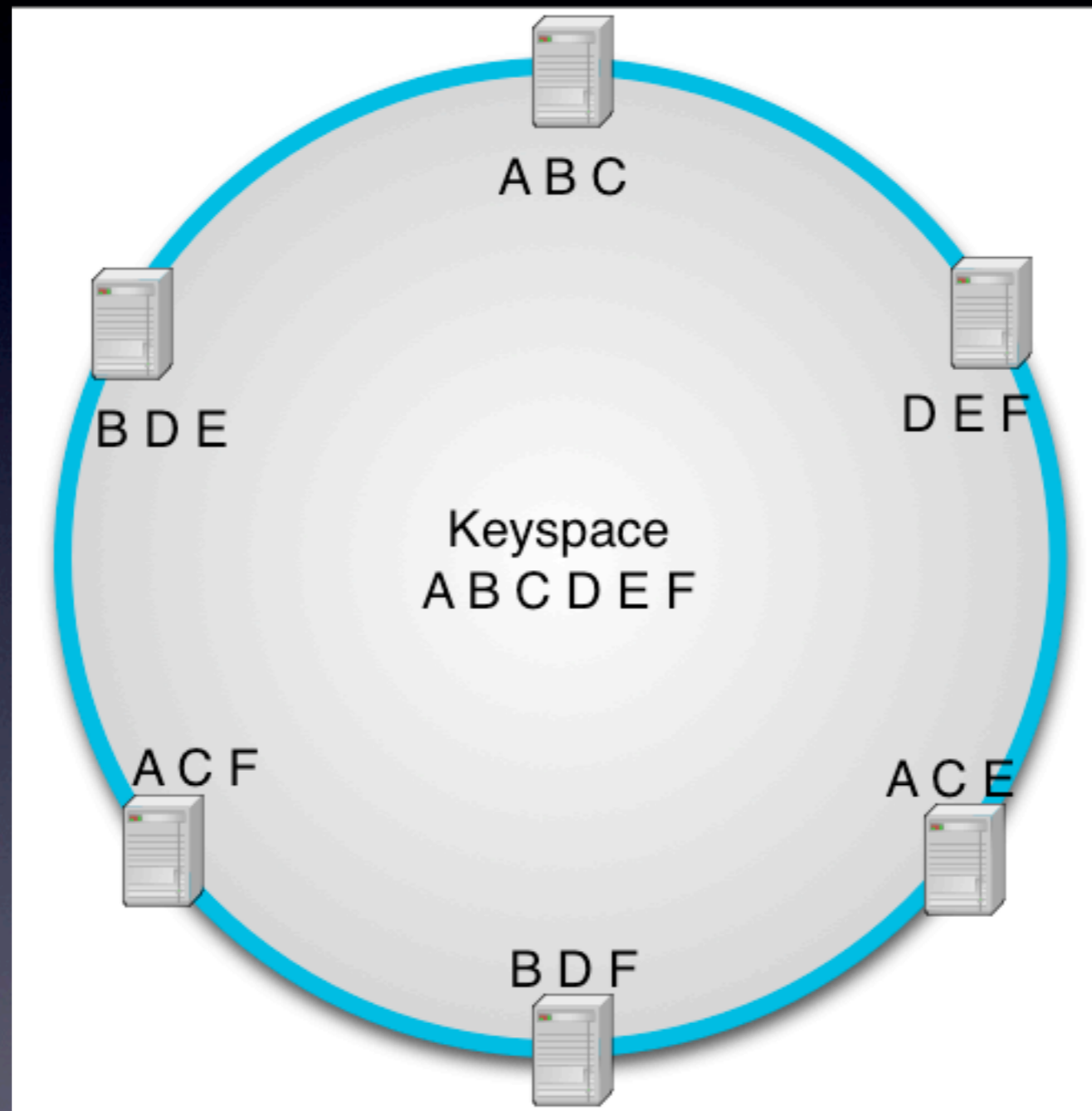


Success | Latent | Short-Circuited | Timeout | Rejected | Failure | Error %



NETFLIX Tweet @jedberg with feedback!

Cassandra



Priam



Priam

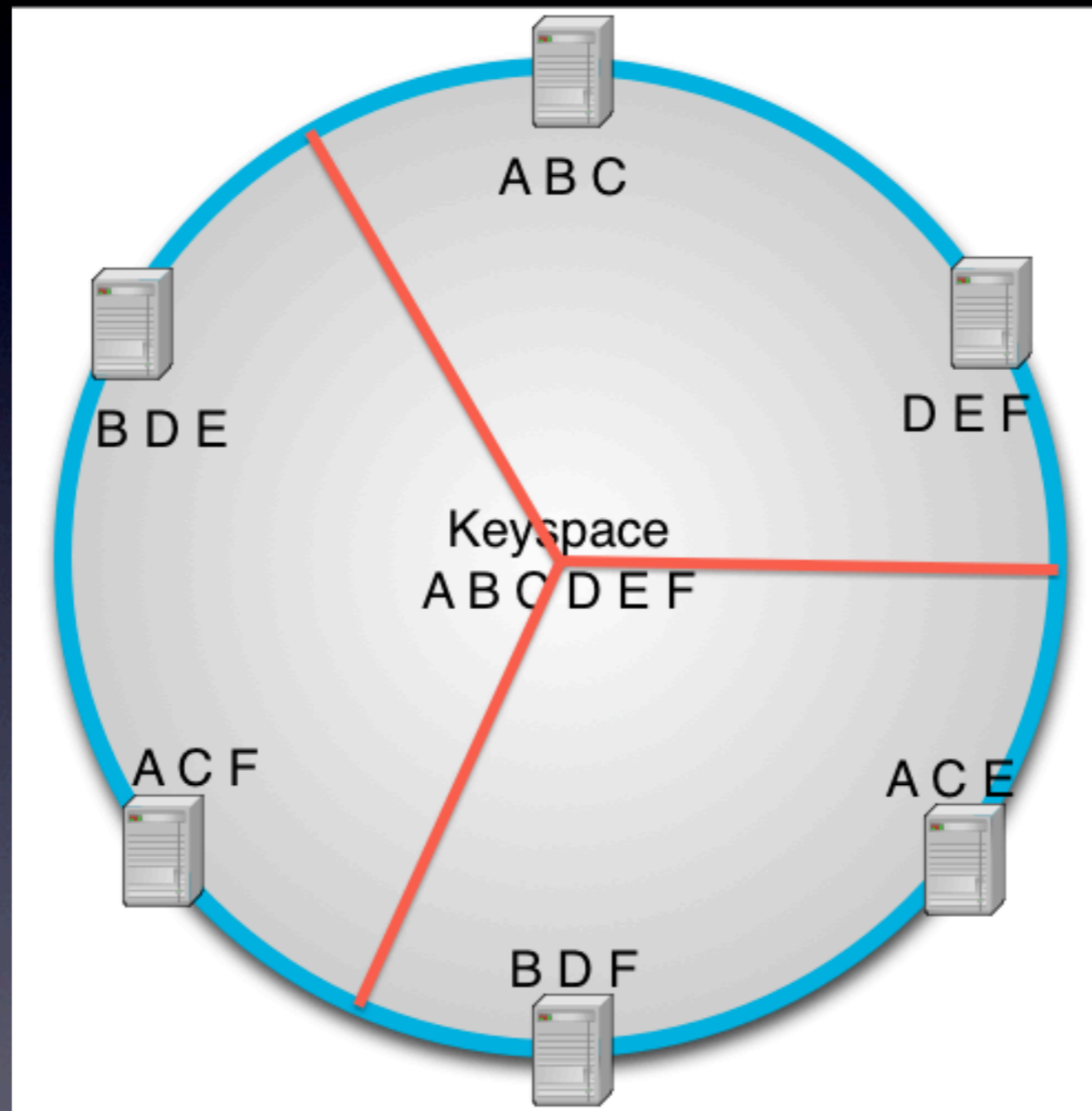
Co-Process for backup/recovery, Token Management, and Centralized Configuration management for Cassandra. [More Info](#)

Watchers:	156
Forks:	40
Language:	Java
Open Issues:	21
Updated:	11/07/12 @21:53:42

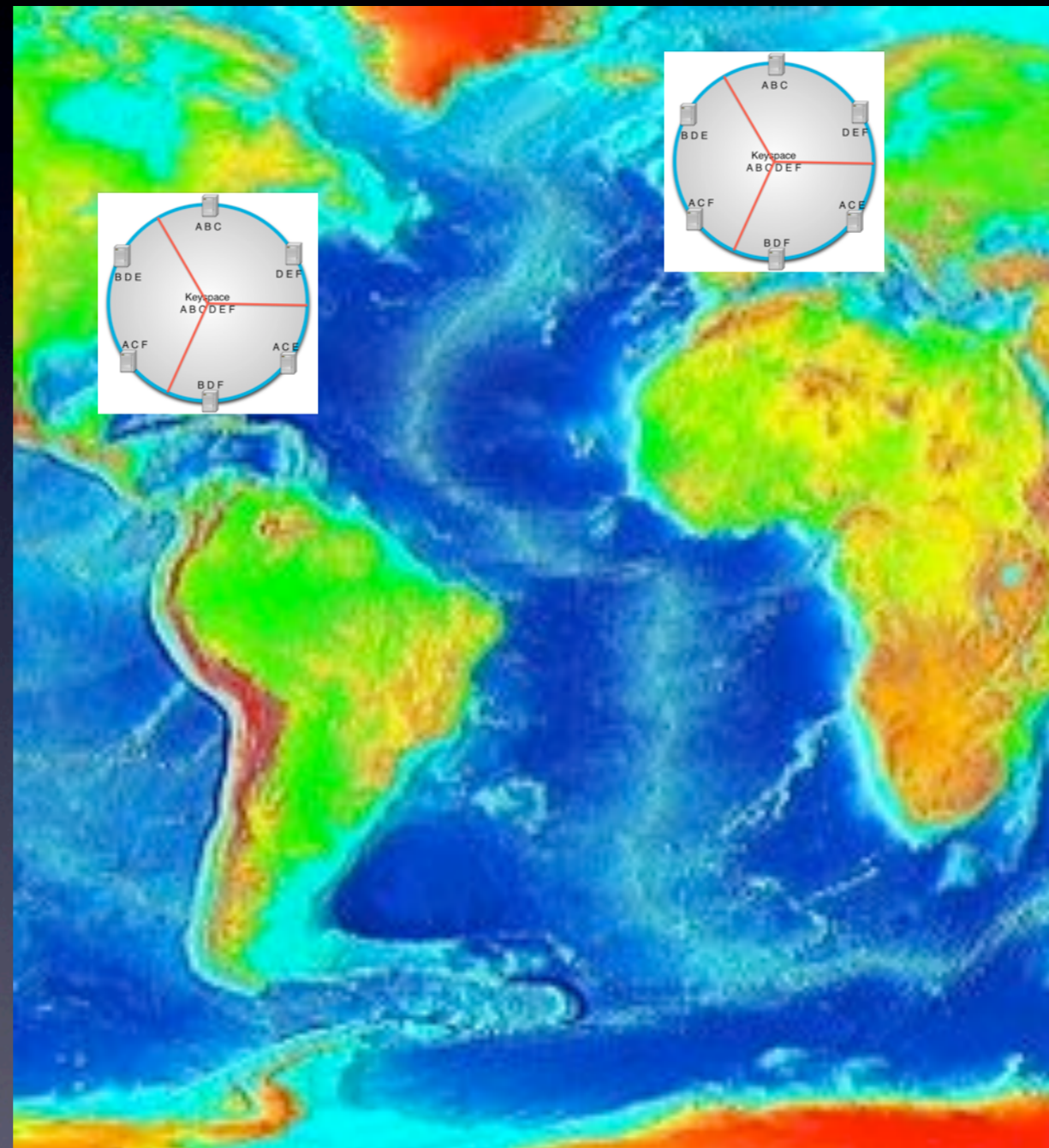


NETFLIX Tweet @jedberg with feedback!

Cassandra Architecture



Cassandra Architecture



NETFLIX Tweet @jedberg with feedback!

How it works

- Replication factor
- Quorum reads / writes
- Bloom Filter for fast negative lookups
- Immutable files for fast writes
- Seed nodes
- Multi-region
- Gossip protocol



Cassandra Benefits

- Fast writes
- Fast negative lookups
- Easy incremental scalability
- Distributed -- No SPoF



Why Cassandra?

- Availability over consistency
- Writes over reads
- We know Java
- Open source + support





NETFLIX Tweet @jedberg with feedback!

We live in an unreliable world



Tweet @jedberg with feedback!



NETFLIX

Tweet [@jedberg](#) with feedback!



NETFLIX Tweet @jedberg with feedback!

MURPHY'S LAW

IF IT CAN GO WRONG,
IT WILL!



NETFLIX Tweet @jedberg with feedback!

Tips, and Tricks



Queues are your friend

- Votes
- Comments
- Thumbnail scraper
- Precomputed queries
- Spam
 - processing
 - corrections



Caching is a good way
to hide your failures

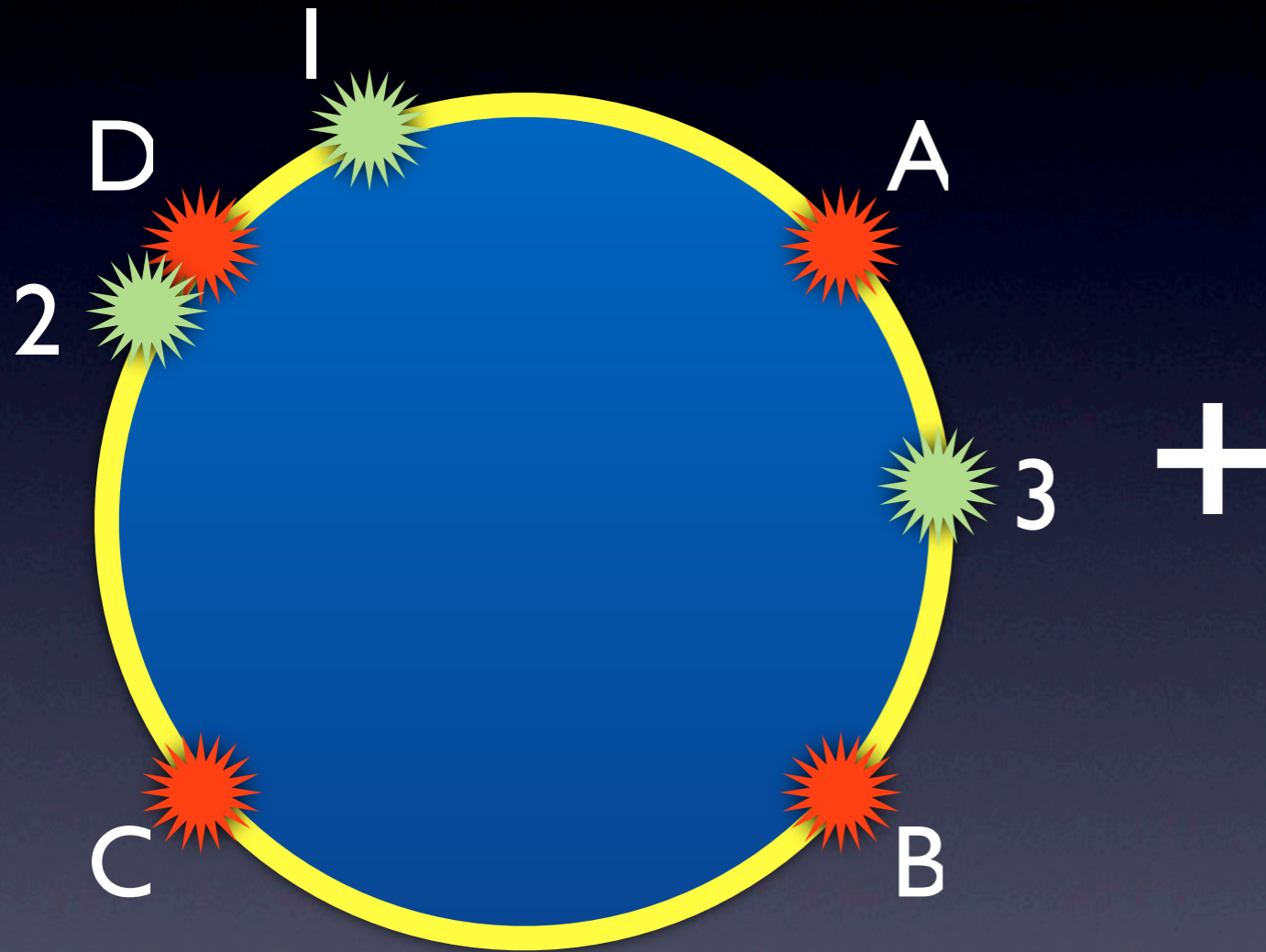


NETFLIX Tweet @jedberg with feedback!

Sometimes users notice your data inconsistency



EVCache



NETFLIX Tweet @jedberg with feedback!

Do you even need a
cache?



NETFLIX Tweet @jedberg with feedback!

Think of SSDs as cheap
RAM, not expensive disk



NETFLIX Tweet @jedberg with feedback!

Going multi-zone or multi-datacenter



NETFLIX Tweet @jedberg with feedback!

Benefits of Amazon's Zones

- Loosely connected
- Low latency between zones
- 99.95% uptime guarantee per zone



Going Multi-region



NETFLIX Tweet @jedberg with feedback!

Leveraging Mutli-region

- 100% uptime is theoretically possible.
- You have to replicate your data
- This will cost money



Other options

- Backup datacenter
- Backup provider



NETFLIX Tweet @jedberg with feedback!

Cause chaos



NETFLIX Tweet @jedberg with feedback!

The Monkey Theory

- Simulate things that go wrong
- Find things that are different



NETFLIX Tweet @jedberg with feedback!

The simian army

- **Chaos -- Kills random instances**
- **Latency -- Slows the network down**
- Conformity -- Looks for outliers
- Doctor -- Looks for passing health checks
- Janitor -- Cleans up unused resources
- Howler -- Yells about bad things



The Chaos Gorilla



NETFLIX Tweet @jedberg with feedback!

Automate all the things!



NETFLIX Tweet @jedberg with feedback!

Automate all the things!

- Application startup
- Configuration
- Code deployment
- **System deployment**



Incident Reviews

Ask the key questions:

- What went wrong?
- How could we have detected it sooner?
- How could we have prevented it?
- How can we prevent this class of problem in the future?
- How can we improve our behavior for next time?



The Netflix way

- Everything is “built for three”
- Fully automated build tools to test and make packages
- Fully automated machine image bakery
- Fully automated image deployment



All systems choices
assume some part will
fail at some point.



NETFLIX Tweet @jedberg with feedback!

Best Practices

- Keep data in multiple Availability Zones / DCs
- Avoid keeping state on a single instance



Best Practices

- Isolated Services
- Three Balanced AZs
- Triple replicated persistence
- Isolated Regions



Best Practices

- Don't trust your dependencies
 - Have good fallbacks
 - Use circuit breakers/dependency commands



Best Practices

- Be generous in what you accept and stingy in what you give



NETFLIX Tweet @jedberg with feedback!

Best Practices

- Hope for the best, assume the worst



NETFLIX Tweet @jedberg with feedback!



© PETER DAM 2009



NETFLIX Tweet @jedberg with feedback!

War Stories



NETFLIX Tweet @jedberg with feedback!

April 20 | EBS outage



NETFLIX Tweet @jedberg with feedback!

June 29th Outage

- Due to a severe storm, power went out in one AZ
- Netflix did not do well because of a bug in our internal mid-tier load balancer
- However, Cassandra held up just fine!



October 29th Outage

- EBS degradation in one Zone
- We did much better this time
- Cassandra just kept running
- MySql not as well, but fallbacks kicked in



Hurricane Sandy

The outage that never was



NETFLIX Tweet @jedberg with feedback!

Just a quick reminder...

(Some of) Netflix is open source:

<https://netflix.github.com/>

NETFLIX Netflix Open Source Center

Repositories Commit Timeline Mailing Lists

Our Repositories

Astyanax Thrillers

Curator ACTION & ADVENTURE

Priam DOCUMENTARY

CassJMeter DRAMA

Servo COMEDY

Aws-Autoscaling Animal TALES

Exhibitor FOREIGN

Archaius SCI-FI & FANTASY

Asgard INDEPENDENT

SimianArmy CLASSICS

Eureka CRIME ACTION

A Netflix Original Production
© 2012 Netflix, Inc. All rights reserved.

Open Source
Netflix Open Source
Netflix GitHub
Mailing Lists
Get in on the fun: [Join Us!](#)

Communication
Our Tech Blog
@NetflixOSS
Slideshare



NETFLIX Tweet @jedberg with feedback!

Another reminder...

reddit is also open source

<https://github.com/reddit>

patches are now being accepted!



NETFLIX Tweet @jedberg with feedback!

Netflix is hiring

<http://jobs.netflix.com/jobs.html>

- or -

email talent@netflix.com and
tell them jedberg sent you



NETFLIX Tweet @jedberg with feedback!

Questions?



NETFLIX Tweet @jedberg with feedback!



NETFLIX Tweet @jedberg with feedback!

Getting in touch

Email: jedberg@gmail.com

Twitter: [@jedberg](https://twitter.com/jedberg)

Web: www.jedberg.net

Facebook: facebook.com/jedberg

Linkedin: www.linkedin.com/in/jedberg

reddit: www.reddit.com/user/jedberg



NETFLIX Tweet [@jedberg](https://twitter.com/jedberg) with feedback!