Perspectives on Cloud Computing

Raghu Ramakrishnan

Yahoo! Fellow

Chief Scientist, Audience and Cloud Computing

(Many slides courtesy of others at Yahoo!)

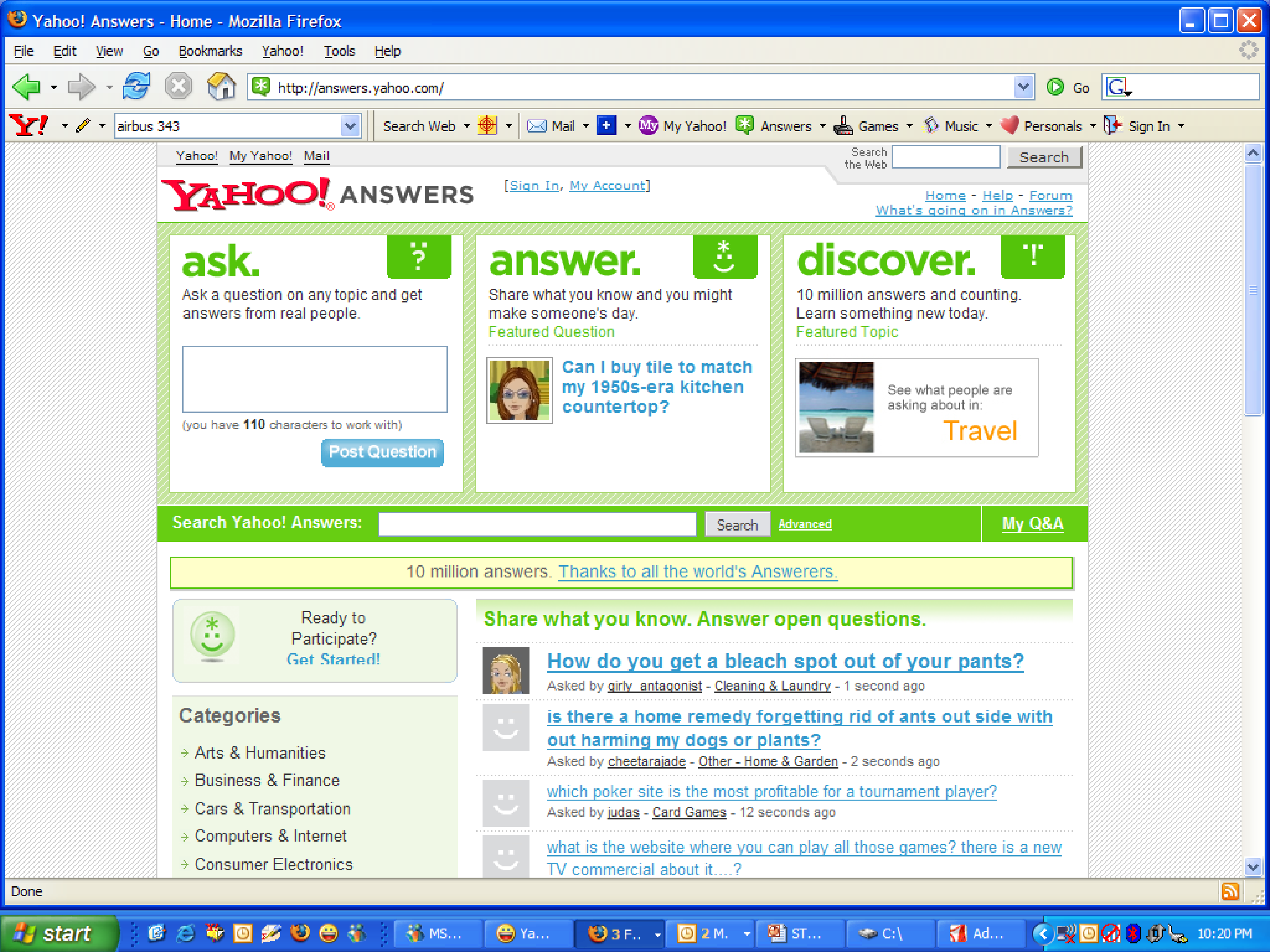
Outline

- Several applications
- Some takeaways on requirements
- Some PNUTS
- Some thoughts on what next (salted throughout the talk)

Requirements for Cloud Services

- **Multitenant.** A cloud service must support multiple, organizationally distant customers.
- **Elasticity.** Tenants should be able to negotiate and receive resources/QoS *on-demand* up to a large scale.
- **Resource Sharing.** Ideally, spare cloud resources should be transparently applied when a tenant's negotiated QoS is insufficient, e.g., due to spikes.
- **Horizontal scaling.** The cloud provider should be able to add cloud capacity in increments without affecting tenants of the service.
- **Metering.** A cloud service must support accounting that reasonably ascribes operational and capital expenditures to each of the tenants of the service.
- **Security.** A cloud service should be secure in that tenants are not made vulnerable because of loopholes in the cloud.
- **Availability.** A cloud service should be highly available.
- **Operability.** A cloud service should be easy to operate, with few operators. Operating costs should scale linearly or better with the capacity of the service.

QUIQ


Yahoo! Answers - Home - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://answers.yahoo.com/
Go
airbus 343
Search Web
Mail
My Yahoo!
Answers
Games
Music
Personals
Sign In
Yahoo! My Yahoo! Mail
Search the Web
Search
YAHOO! ANSWERS
[Sign In, My Account]
Home - Help - Forum
What's going on in Answers?
ask.
Ask a question on any topic and get answers from real people.
(you have 110 characters to work with)
Post Question
answer.
Share what you know and you might make someone's day.
Featured Question
Can I buy tile to match my 1950s-era kitchen countertop?
discover.
10 million answers and counting. Learn something new today.
Featured Topic
See what people are asking about in:
Travel
Search Yahoo! Answers:
Search
Advanced
My Q&A
10 million answers. Thanks to all the world's Answerers.
Ready to Participate?
Get Started!
Categories
Arts & Humanities
Business & Finance
Cars & Transportation
Computers & Internet
Consumer Electronics
Share what you know. Answer open questions.
How do you get a bleach spot out of your pants?
Asked by girly_antagonist - Cleaning & Laundry - 1 second ago
is there a home remedy forgetting rid of ants out side with out harming my dogs or plants?
Asked by cheetarajade - Other - Home & Garden - 2 seconds ago
which poker site is the most profitable for a tournament player?
Asked by judas - Card Games - 12 seconds ago
what is the website where you can play all those games? there is a new TV commercial about it....?
Done
start
MS...
Ya...
3 F..
2 M.
ST...
C:\
Ad...
10:20 PM

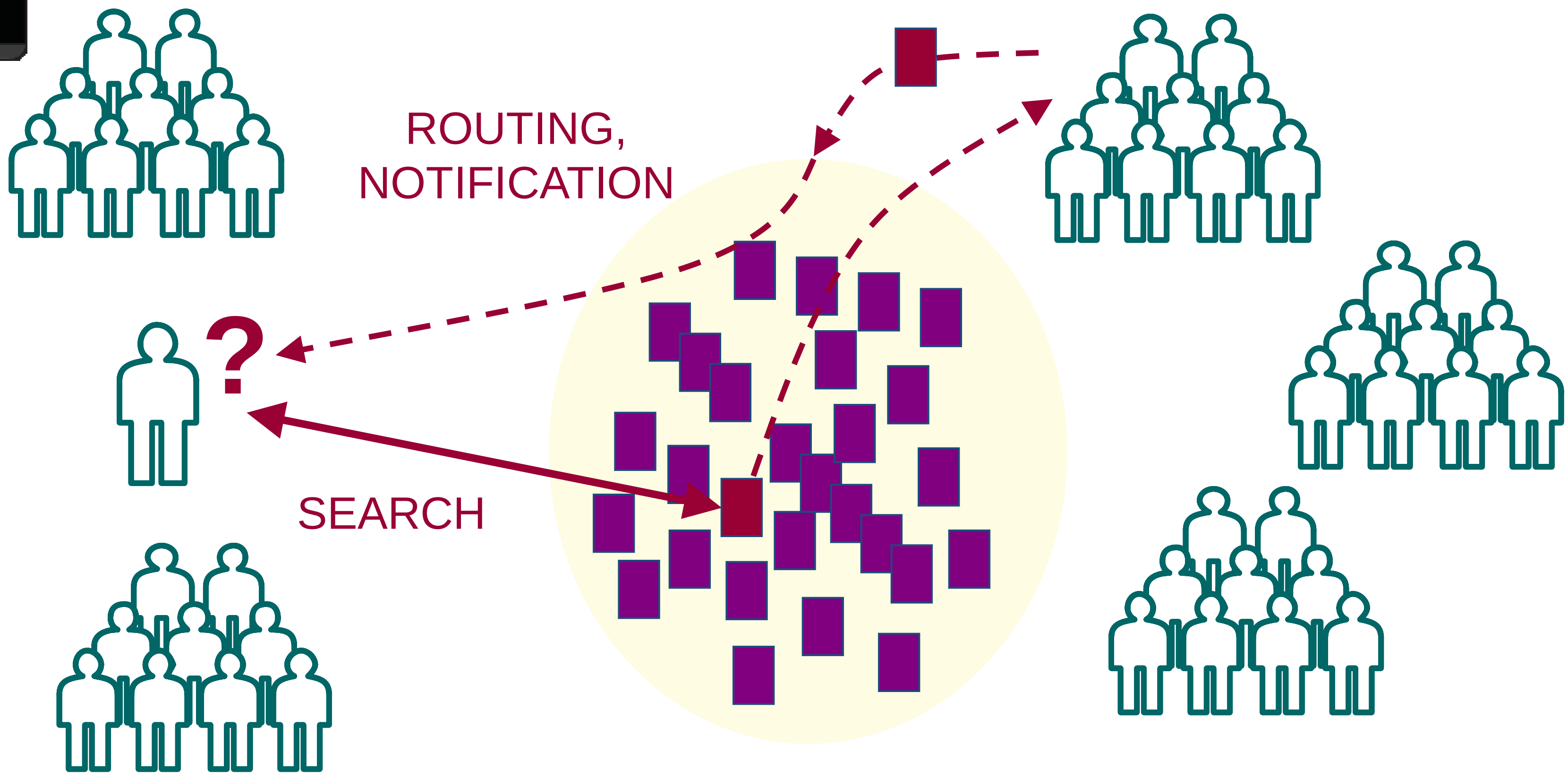
TECH SUPPORT AT COMPAQ

"In newsgroups, conversations disappear and you have to ask the same question over and over again. The thing that makes the real difference is the ability for customers to collaborate and have information be persistent. That's how we found QUIQ. It's exactly the philosophy we're looking for."

"Tech support people can't keep up with generating content and are not experts on how to effectively utilize the product … Mass Collaboration is the next step in Customer Service."

– Steve Young, VP of Customer Care, Compaq

MASS COLLABORATION FOR CRM
aka Crowd-Sourcing

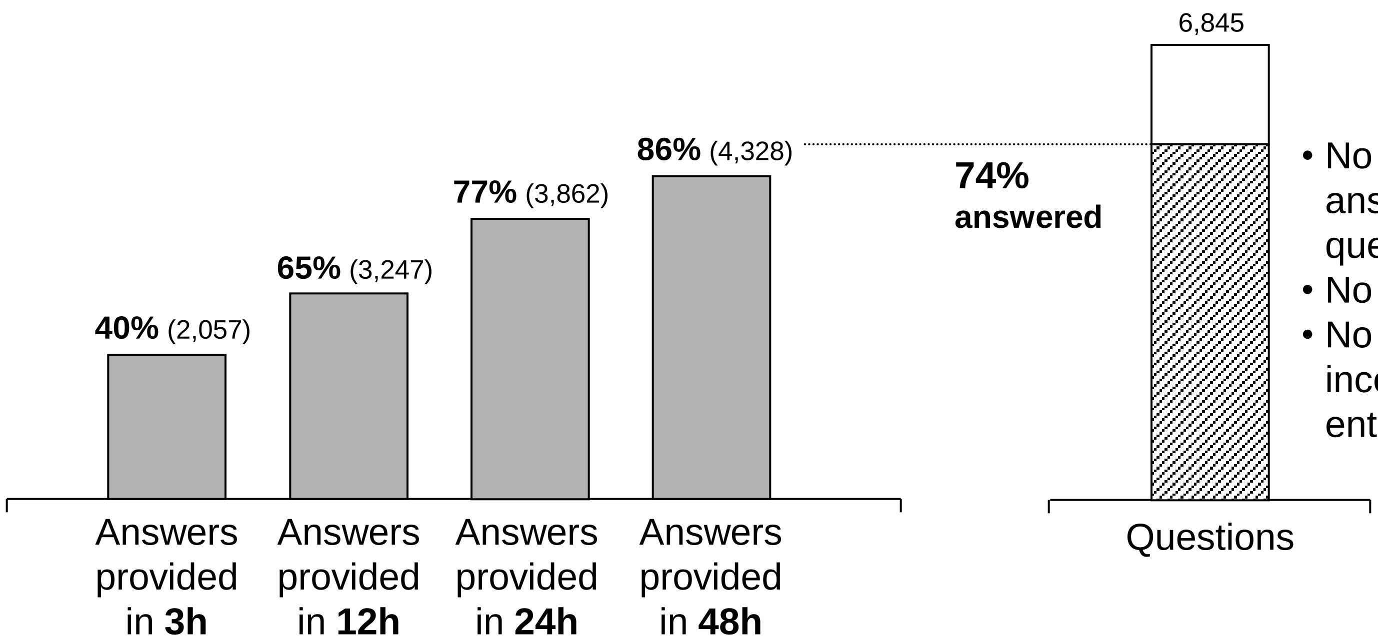


“If it’s not there, find someone who knows”
And make “it” easy to find later

TIMELY ANSWERS

▶ 77% of answers are provided within 24h

But are any good answers? Yes, but most are bad.



- No effort to answer each question
- No added experts
- No monetary incentives for enthusiasts

Answer quality, trust, reputation

SaaS Multitenancy

Long tail of tenants with same logical database scheme

custid	qid	hierarchy	qhdr	qtxt	asker	#ans	about
compaq	22	Sup/presario/ security	“How do I …”	… more details …	Bill Gates	7	“kournikova virus …”

Handling growth

- Small customers multitenant a single table instance
- As they grow, large customers spilled into their own table instance
  - Even with same indexes etc., very different data distributions
  - Want (and will pay for) different SLAs, isolation, audit trails …

Non-Serializable Transactions

Asynchronously updated count
De-normalized design, btw

custid	qid	hierarchy	qhdr	qtxt	asker	#ans	about
compaq	22	Sup/presario/ security	“How do I …”	… more details …	Bill Gates	7	“kournikova virus …”

What’s good, Phaedrus?

- App designers use these tricks all the time, gaining performance by leveraging some semantic slack
  - Though life can get messy if the developer loses track of all the assumptions about what’s acceptable …

Batch-Updates to Online Table

These fields periodically updated by a privileged user
All rows are affected

custid	qid	hierarchy	qhdr	qtxt	asker	#ans	about
compaq	22	Sup/presario/ security	"How do I …"	… more details …	Bill Gates	7	"kournikova virus …"

Two flavors—Atomic & Not

- Hierarchy:
  - Changes across all rows must appear atomic
- About:
  - Rows can be updated one at a time

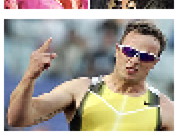
Could have replaced RDBMS by key-value store for this app!
(We needed indexes, but built our own outside DBMS anyway)

COKE

Cloud Services @Y!: Use Cases

Today Module

Product Objective	Prioritize small pool of editorially programmed packages to optimize engagement in real-time

Package	Young Adults (2)	Social Chairmen (2)	CHO (2)	Young Boomers (2)	Older Boomers (2)
All Below Packages	2,900,789 32,086 1.11 -35.83	4,470,060 58,201 1.30 -24.47	3,537,594 38,777 1.10 -36.41	1,781,139 18,966 1.06 -38.23	2,813,807 27,105 0.96 -44.12
	6 49,400 788 1.60 0.07	7 74,808 1,214 1.62 1.80	1 94,917 1,929 2.03 27.49	1 26,511 614 2.32 45.29	3 32,043 586 1.83 14.72
	4 26,703 607 2.27 14.45	6 57,930 1,230 2.12 6.90	2 22,539 551 2.44 23.08	2 12,666 353 2.79 40.32	1 15,426 414 2.68 35.12

Key Features

Package Ranker (COKE)

Ranks packages by expected CTR based on data collected every 5 minutes

Dashboard (COKE)

Provides real-time insights into performance by package, segment, and property

Mix Management (Property)

Ensures editorial voice is maintained and user gets a variety of content

Package rotation (Property)

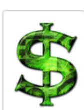
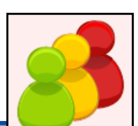
Tracks which stories a user has seen and rotates them after user has seen them for a certain period of time

Key Performance Indicators

160% Lift in CTR

Editorial Voice Preserved

Approaches

	Estimate Most Popular (EMP)	*"What's most engaging overall?"*
	Behavioral Affinities	*"People who did X, did Y"*
Italian RED 94087 94089 RED Italian	Attribute Similarities	*"Related items with similar metadata"*
	Business Optimization	*"What generates most business value?"*
	Personalized Recommendations	*"What's most relevant to <u>me</u> based on my interests and attributes?"*
	Social Recommendations	*"What are my trusted connections into?"*

EMP Challenges

Highly dynamic system:

- Short article lifetimes
- Pool constantly changing
- User population is dynamic
- CTRs non-stationary

Content Optimization Overview

Online Learning

- Online regression models, time-series models
- Model the temporal dynamics
- Provide fast learning for per-item models

Offline Modeling

- Exploratory data analysis
- Regression, feature selection, collaborative filtering (factorization)
- Seed online models & explore/exploit methods at good initial points
- Reduce the set of candidate items

Near real-time user feedback

Explore/Exploit

- Multi-armed bandits
- Find the best way of collecting real-time user feedback (for new items)

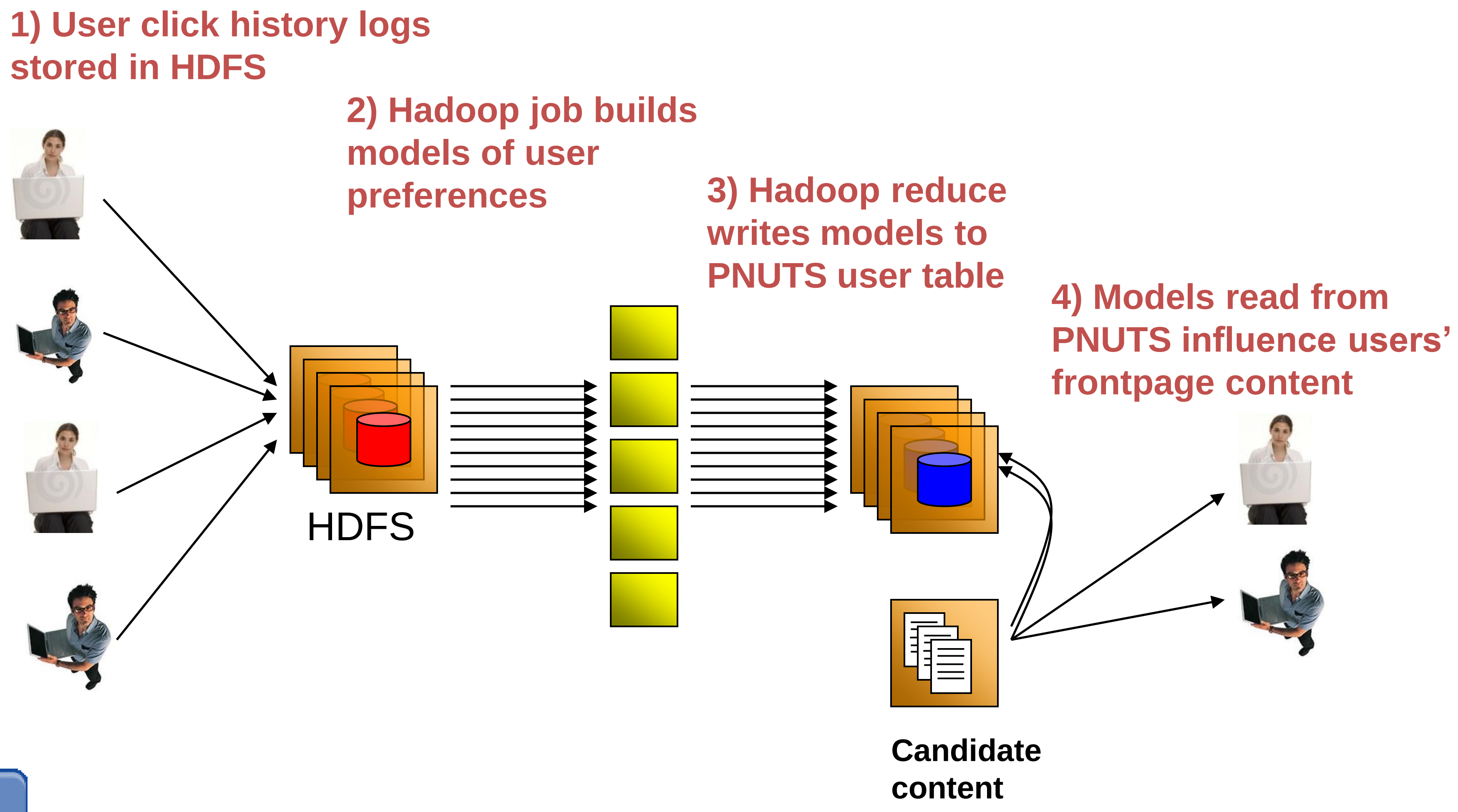
Large amount of historical data (user event streams)

Data pipelines

PNUTS

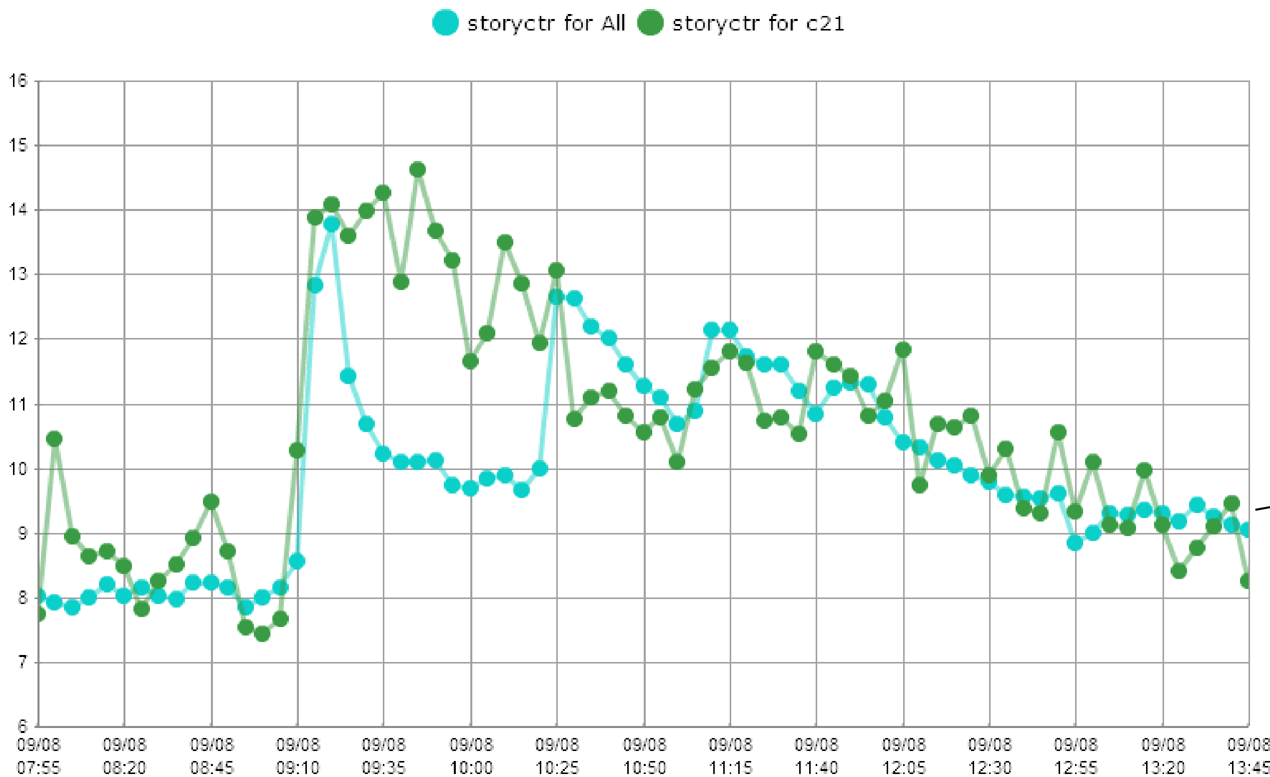
Hadoop

Data Management in COKE

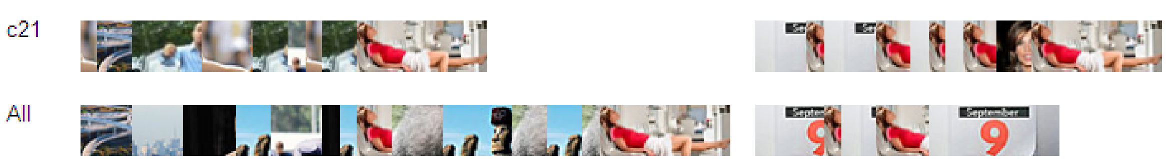


COKE Dashboard: Overall CTR

Compare performance of models and historical benchmarks



Compare buckets and models over time



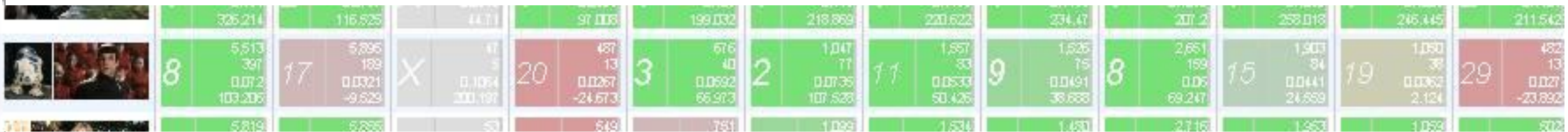
See which content was promoted most across time

Bucket	Page Views,	All Clicks	Story Clicks	Footer Clicks	OverAll CTR	Story CTR	Footer CTR	Lift compared to All
All	43,005,783	13,061,688	4,286,407	8,775,281	30.37	9.97	20.4	0
c21	281,821	84,853	29,777	55,076	30.11	10.57	19.54	6.02

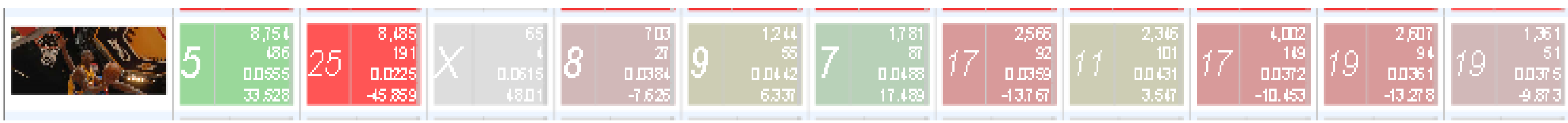
Compare bucket metrics

Examples

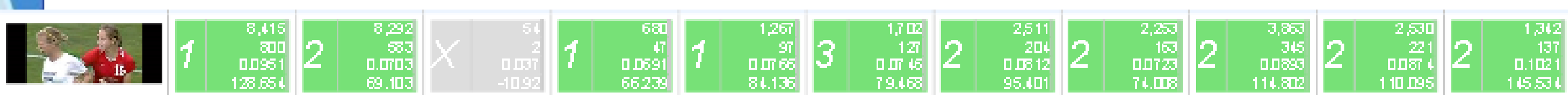
- **ACQUISITION:** A “Star Trek” package was #3 with 18-20 demo, #2 with 21-24 demo, but #9 overall. We can acquire younger audiences with targeted content like this.



- **ENGAGEMENT:** “Kobe’s astonishing shot” was #25 with women, but #5 with men. We can better engage men (or sports fans) by showing more like this, women by showing less.



- **REACH:** A package about a hair-pulling soccer player was just plain interesting to everyone (#1-3). We can maintain reach by programming content for the mass audience.



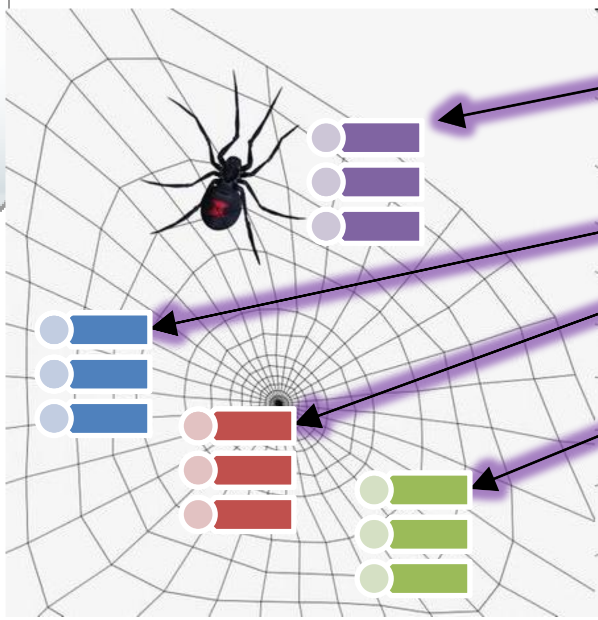
WoC


Concept-centric
Information aggregation
berts - Yahoo! Search Results - Mozilla Firefox
View History Bookmarks Yahoo! Tools Help
http://search.yahoo.com/search?p=julia+roberts&ei=UTF-8&fr=moz35
r sunnyvale - Yahoo! Search Re...
julia roberts - Yahoo! Search Res...
YAHOO!
julia roberts
Search
julia roberts twins
julia roberts movies
lyle julia roberts
julia roberts babies
julia roberts henry daniel moder
Explore related concepts:
actor
julia roberts photos
Julia Roberts News
Search Pad
SearchScan - On
40,500,000 results for
julia roberts
Related People
Scarlett Johansson
Lindsay Lohan
Tom Hanks
Halle Berry
Julia Roberts - Image Results
News & Photos
Twitter
Hindus concerned about Julia Roberts' "Eat, Pray, Love" - New Kerala - 6 hours ago
Trailer For 'Eat Pray Love' Starring Julia Roberts - KPBS San Diego - Mar 19 03:18pm
Link Party: Julia Roberts' New Movie Will Teach You How to Live - E! Online - Mar 18 05:48pm
more Julia Roberts news...
Julia Roberts - Wikipedia
Early life | Career | Influence | Personal life
Julia Fiona Roberts is an American actress. She is known for starring in the romantic comedy Pretty Woman opposite Richard Gere, which grossed $463 million worldwide. After receiving...
en.wikipedia.org/wiki/Julia_Roberts - 122k - Cached

Web of Concepts

rich, aggregated data

concept



madonna

mumbai restaurant

san jose

• • •

Aggregated KB

INDEX

SERP

The "index" is keyed by concept instance, and organizes all relevant information (data describing the concept instance and its relationship to other instances), wherever it is drawn from, in semantically meaningful ways santa clara - Yahoo! Search Results - Mozilla Firefox

History Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search?p=shower+head+santa+clara&ei=UTF-8&fr=moz35

shower head santa clara

vale - Yahoo! Search Re... | shower head santa clara - Yahoo!...

YAHOO!

Web Images Video Local Shopping More

shower head santa clara

Search

Options

Start typing to see suggestions.

Explore related concepts:

pool | tile
Detached Single Family | MLSListings
Tub | toilets
faucets | Real Estate MLS

Settings

Search Pad

SearchScan - On

577,000 results for
shower head santa cl...:

Show All

Google Sites

Yahoo! Local

Metacafe

Video Sites

Sponsored Results

Shower Head
High Performance Showerheads for Full Body Coverage. Learn More.
Moen.com/Showerheads

Shower Head stores near Santa Clara, CA

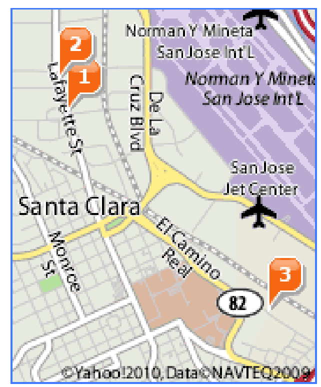
Nearby City

All (26)	San Jose (8)	Mountain View (3)	Los Gatos (2)
Santa Clara (2)	Fremont (3)	Campbell (3)	Sunnyvale (1)

1 Conleff Plumbing Supply ★★★★★ (5)
conleff.com
(408) 988-8005 - 2301 Lafayette St, **Santa Clara**, CA
5 Reviews | Overview | 1 Photo | Directions

2 Home Depot ★★★★★ (9)
homedepot.com
(408) 492-9600 - 2435 Lafayette St, **Santa Clara**, CA
5 Reviews | Overview | Directions

3 Kitchen & Bath Showplace ★★★★★ (1)
kbshowplace.com
(408) 249-9880 - 1200 Campbell Ave, San Jose, CA
Overview | 1 Photo | Directions



23 More Local Results...

Next Previous Highlight all Match case Reached end of page, continued from top

2 Windows... | shower hea... | Inbox - Micr... | Microsoft Po... | Search Desktop

ry Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search;_ylt=A0oGklHEA_RLAyYBgxJXNyoA?p=eggplant+parmigiana+baltimore&fr2=sb-top&fr=moz35&sao=1

na baltimore - Yah...

YAHOO!

Web Images Video Local Shopping News More

eggplant parmigiana baltimore

Search Options

Start typing to see suggestions.

Explore related concepts:

Pizza
veal
Chicken
Italian Restaurant
mushrooms
tomato sauce
Little Italy
Crab Cake

Search Pad

SearchScan - On

30,300 results for **eggplant parmigiana ...:**

Show All

Los Angeles Times

Local Business Sites

Eggplant Parmigiana Restaurants near Baltimore, MD

Neighborhood

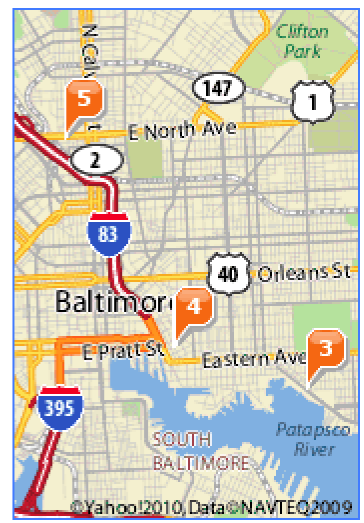
All (36) Canton (2) Charles North (1) Downtown (1)
Abell (1) Central Bal... (4) Chinquapin ... (1) Federal Hill (1)

1 Ciao Bella - **Baltimore** ★★★★★ (11)
local.yahoo.com
(410) 685-7733 - 236 S High St, **Baltimore**, MD
Menu: **eggplant parmigiana**
4 Reviews | Overview | 2 Photos | Directions

2 Amicci's ★★★★★ (20)
amiccis.com
(410) 528-1096 - 231 S High St, **Baltimore**, MD
Menu: **eggplant parmigiana**
14 Reviews | Overview | 23 Photos | Directions

3 Pasticcio ★★★★ (8)
local.yahoo.com
(410) 522-7700 - 2400 Boston St, **Baltimore**, MD
Menu: **eggplant parmigiana**
5 Reviews | Overview | 3 Photos | Directions

4 Caesar's Den ★★★★ (7)
caesarsden.com
(410) 547-0820 - 223 S High St, **Baltimore**, MD
Menu: **eggplant parmigiana**
4 Reviews | Overview | 11 Photos | Directions



Search Meets Structured Data

Searches (often) retrieve data from tables

- Can pre-compute tables/indexes and push to serving tier periodically
    - Batch updates in an extreme sense
- Want to be able to scale (read-only) serving system as effectively as traditional IR based infrastructure

Hadoop Core (Core, Pig, Oozie, Hive, Howl)	Ad BT and Inventory prediction, Content Agility, UDA, COKE, Mail Spam, Search, APG, Labs, Insights, Analytics	1+ million jobs per month 3.7 PB processed daily 90B events and 120 TB daily 70+ PB of Data

Map-Reduce and more …

HADOOP: SCALABLE ANALYTICS

Abuse/Spam Overview

Raw Data

- UGC
  Email, IM, Answers
- Activity
  User log, clicks
- User Profile
  Geo, Age, Gender, Usage pattern

Derived Data

- Abuse Profile
  Yuid, IP, bcookie
- Social Network
  Friend, Community
- Ground Truth

Technology

- Content Classification
- Abnormal Event Detection
- User Disambiguation
- Data Normalization
- Social Relevance

Detection

- Mail Spam
  Content-Based Classifier
- IP Reputation
  Content-Based Classifier

Monitoring

- Traffic Protection
  Abnormal click/view detection
- Unique Users
  User Disambiguation

Prevention

- Trusted Users
  Reputation classifier
- CAPTCHAs

Application: Mail Spam Filtering

Scale of the problem

- ~ 25B Connections, 5B deliveries per day
- ~ 450M mailboxes

User feedback on spam is often late, noisy and not always actionable

Problem	Algorithm	Data size	Running time on Hadoop
Detecting spam campaigns	Frequent Itemset mining	~ 20 MM spam votes	1 hour
“Gaming” of spam IP votes by spammers	Connected component (squaring a bi-partite graph)	~ 500K spammers, 500k spam IPs	1 hour

Example: User Activity Modeling

Large dimensionality vector describing possible user activities
But a typical user has a sparse activity vector

Attribute	Possible Values	Typical values per user
Pages	~ MM	10 – 100
Queries	~ 100s of MM	Few
Ads	~ 100s of thousands	10s

Hadoop pipeline to model user interests from activities

1a. Data Acquisition

Input

- Multiple user event feeds (browsing activities, search, etc.) per time period

User	Time	Event	Source
U_1	T_0	visited autos.yahoo.com	Web server logs
U_1	T_1	searched for “car insurance”	Search logs
U_1	T_2	browsed stock quotes	Web server logs
U_1	T_3	saw an ad for “discount brokerage”, but did not click	Ad logs
U_1	T_4	checked Yahoo Mail	Web server logs
U_1	T_5	clicked on an ad for “auto insurance”	Ad logs, click server logs

1a. Data Acquisition

Output:

- Single normalized feed containing all events for all users per time period

User	Time	Event	Tag
U_1	T_0	Content browsing	Autos, Mercedes Benz
U_2	T_2	Search query	Category: Auto Insurance
…	…	…….	………
...	…	…….	………
U_{23}	T_{23}	Mail usage	Drop event
U_{36}	T_{36}	Ad click	Category: Auto Insurance

1b. Feature and Target Generation

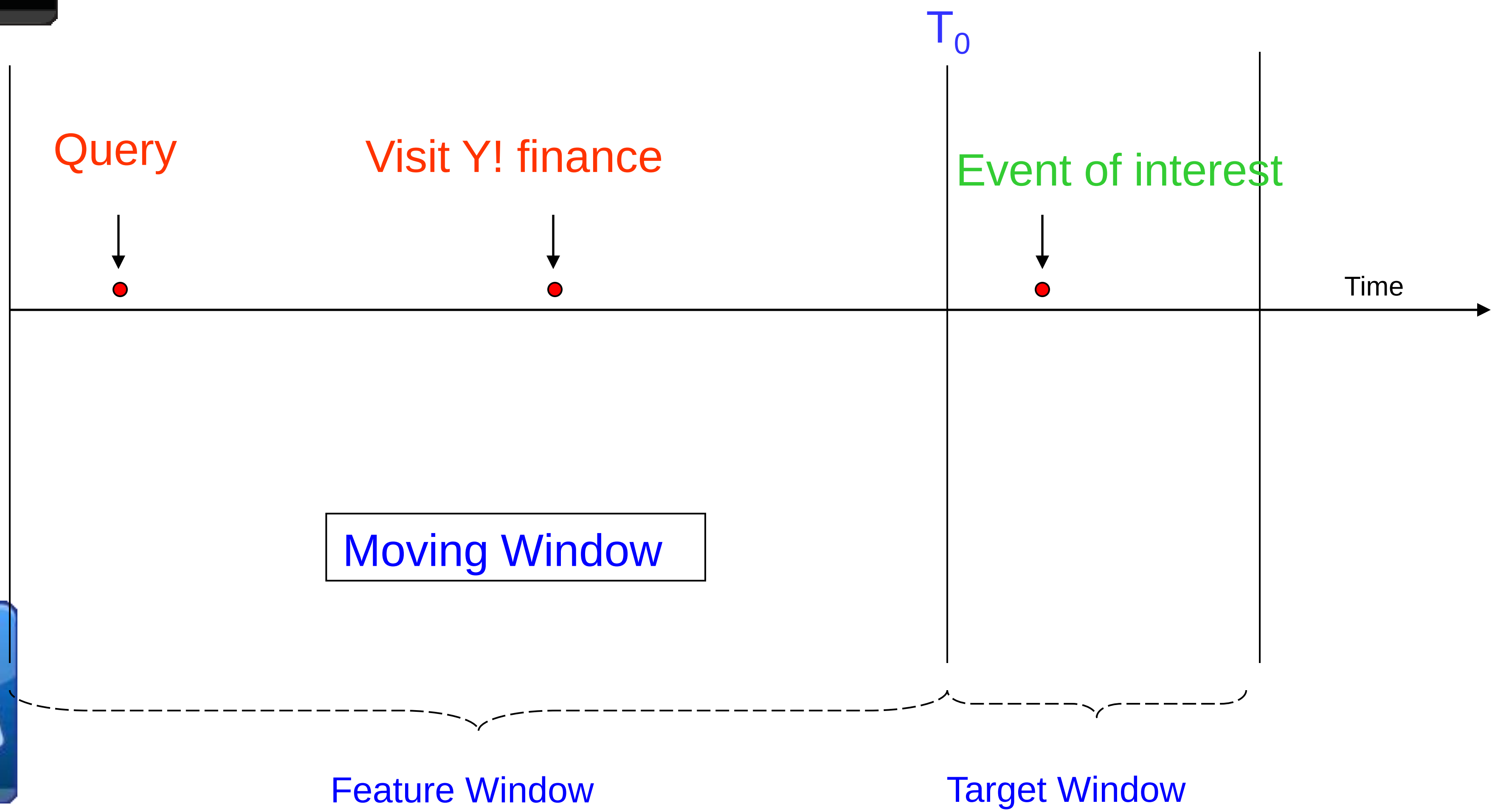
Features:

- Summaries of user activities over a time window
- Aggregates, Moving Averages, Rates, etc., over moving time windows
- Support online updates to existing features

Targets:

- Constructed in the offline model training phase
- Typically, user actions in the future time period indicating interest
  - Clicks/Click-through rates on ads and content
  - Site and page visits
  - Conversion events
    - Purchases, Quote requests etc.
    - Sign-ups to newsletters, Registrations etc.

1b. Feature and Target Windows



34

User Modeling Pipeline

Component	Data Processed	Time
Data Acquisition	~ 1 Tb per time period	2 – 3 hours
Feature and Target Generation	~ 1 Tb * Size of feature window	4 - 6 hours
Model Training	~ 50 - 100 Gb	1 – 2 hours for 100's of models
Scoring	~ 500 Gb	1 hour

Hadoop Pipelines

- Pipeline workflows run repeatedly (e.g., daily, hourly)
- Incremental evaluation support needed
  - Semi-naïve style techniques can help
  - NOVA and other projects
- Soft real-time constraints
- Natural point to inject streaming analytics
- Key observation—Hadoop is being used as more than an analytics platform!
  - Data acquisition, warehouse
  - Lots to optimize here—e.g., # copies of shared files

Renting vs. buying, and being DBA to the world …

DATA MANAGEMENT IN THE CLOUD

Yahoo! Data: Unprecedented Scale

Massive user base and engagement

- 500M+ unique users per month
- Hundreds of petabytes of storage
- Hundreds of billions of objects
- Hundreds of thousands of requests/sec

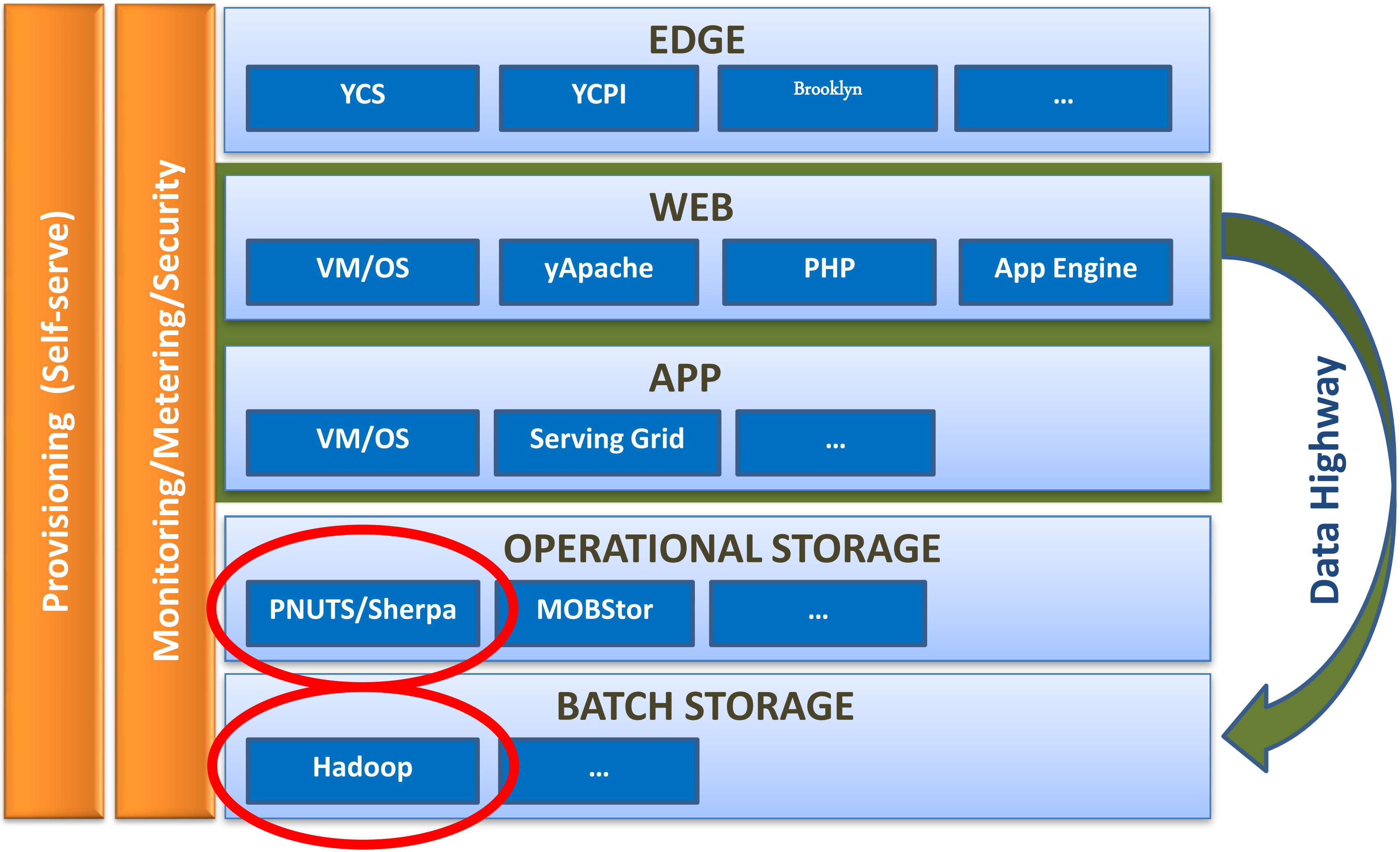
Global

- Tens of globally distributed data centers
- Serving each region at low latencies

Challenging Users

- Rapidly extracting value from voluminous data
- Downtime is not an option (outages cost $millions)
- Variable usage patterns

Yahoo! Cloud Stack



Y!OS, COKE, LocDrop, Video, Media
Search history, Answers, Messenger,
BOSS, Image Search, Blog Search

15K requests per second
Over 1.5B records; 10sTB of data

ACID or BASE? Litmus tests are colorful, but the picture is cloudy

PNUTS:
SCALABLE DATA SERVING

Typical Y! Applications

User logins and profiles

- Including changes that must not be lost!
  - But single-record "transactions" suffice

Events

- Alerts (e.g., news, price drops)
- Social network activity (e.g., user goes offline)
- Ad clicks, article clicks

Application-specific data

- Postings in message board
- Uploaded photos, tags
- Shopping carts

What is PNUTS/Sherpa?

A	42342	E
B	42521	W
C	66354	W
D	12352	E
E	75656	C
F	15677	E

Parallel database

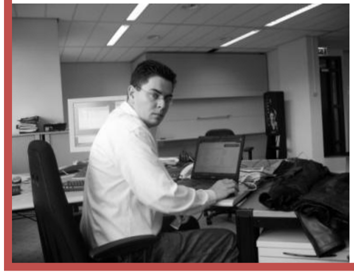
```
CREATE TABLE Parts (
    ID VARCHAR,
    StockNumber INT,
    Status VARCHAR
    …
)
```

Structured, flexible schema

A	42342	E
B	42521	W
C	66354	W
D	12352	E
E	75656	C
F	15677	E

Geographic replication

A	42342	E
B	42521	W
C	66354	W
D	12352	E
E	75656	C
F	15677	E

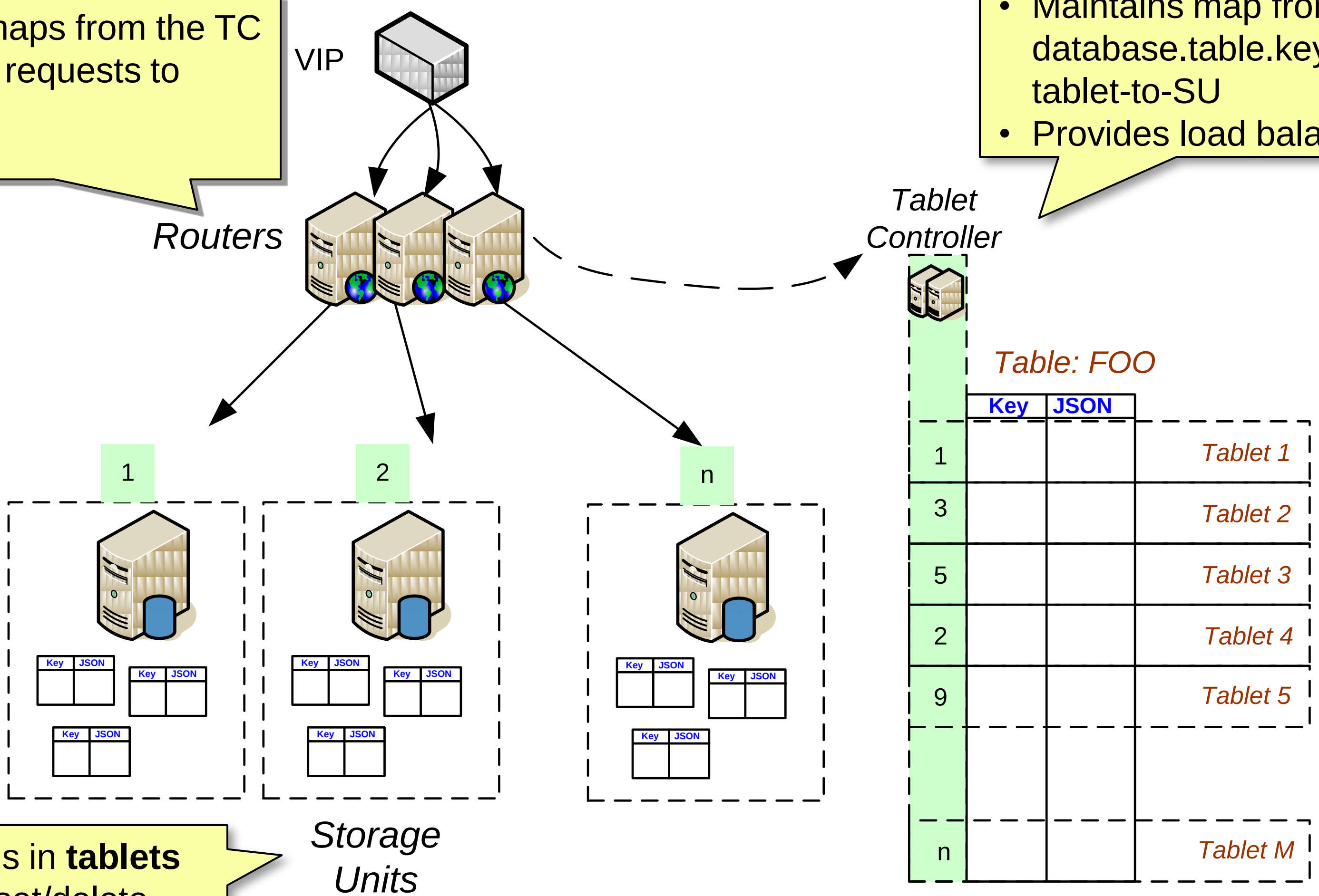


Hosted, managed infrastructure

PNUTS: Key Components

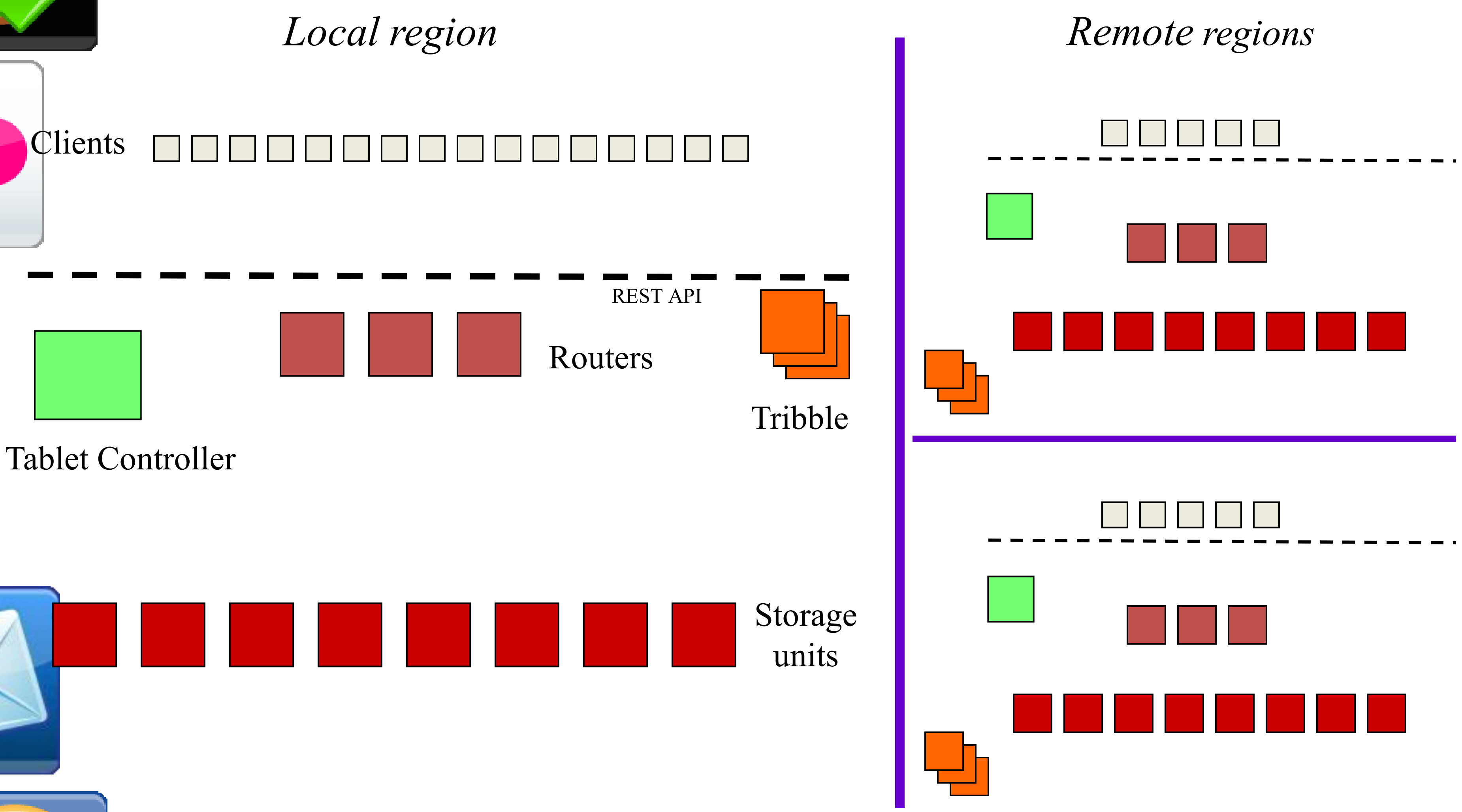
- Caches the maps from the TC
- Routes client requests to correct SU

- Maintains map from database.table.key-to-tablet-to-SU
- Provides load balancing



- Stores records in **tablets**
- Services get/set/delete requests

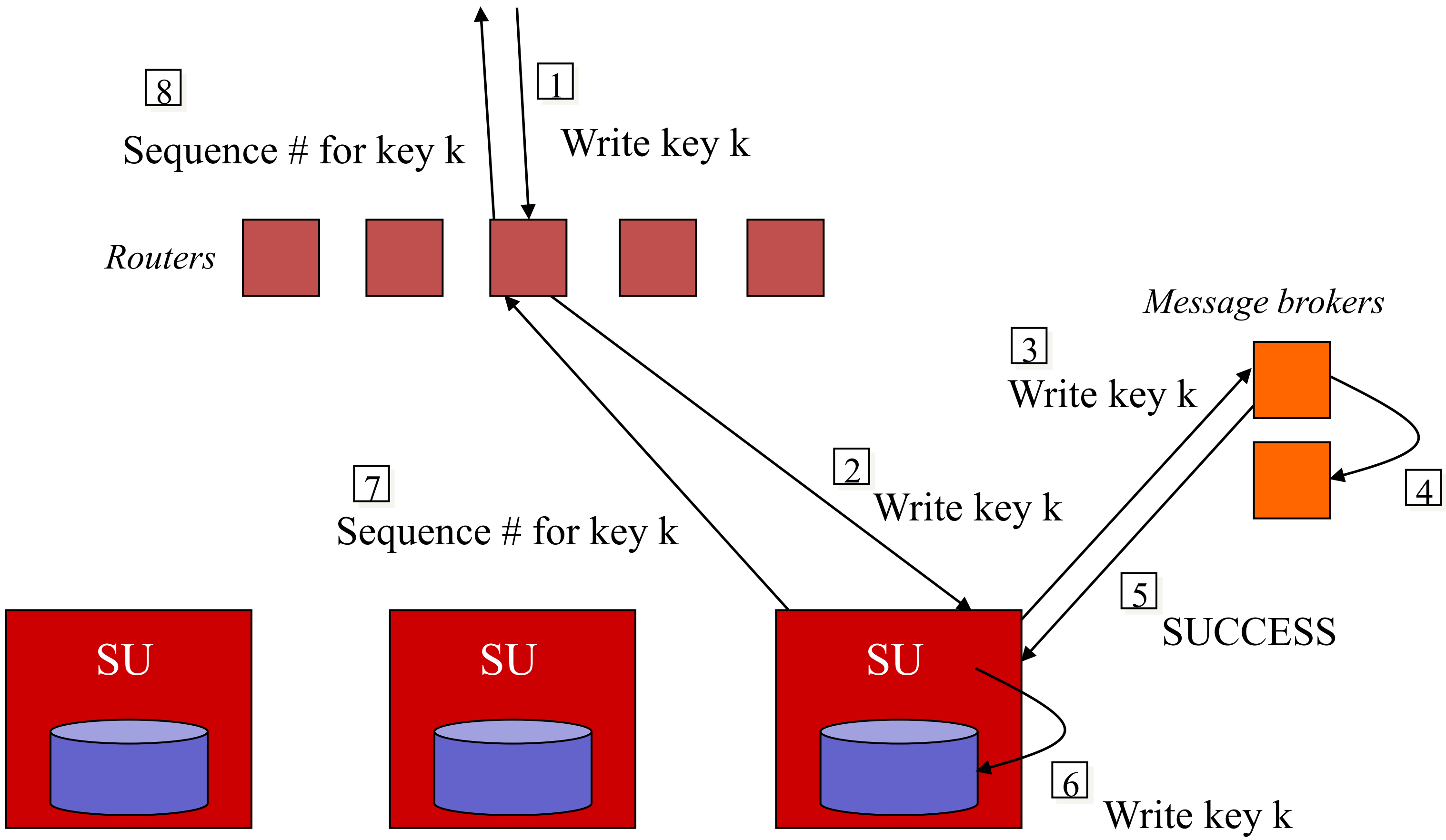
Architecture



Flexible Schema

Posted date	*Listing id*	*Item*	*Price*	*Color*	*Condition*
6/1/07	424252	Couch	$570		Good
6/1/07	763245	Bike	$86		
6/3/07	211242	Car	$1123	Red	Fair
6/5/07	421133	Lamp	$15		

Updates


8
Sequence # for key k
1
Write key k
Routers
Message brokers
3
Write key k
2
Write key k
4
7
Sequence # for key k
5
SUCCESS
SU
SU
SU
6
Write key k

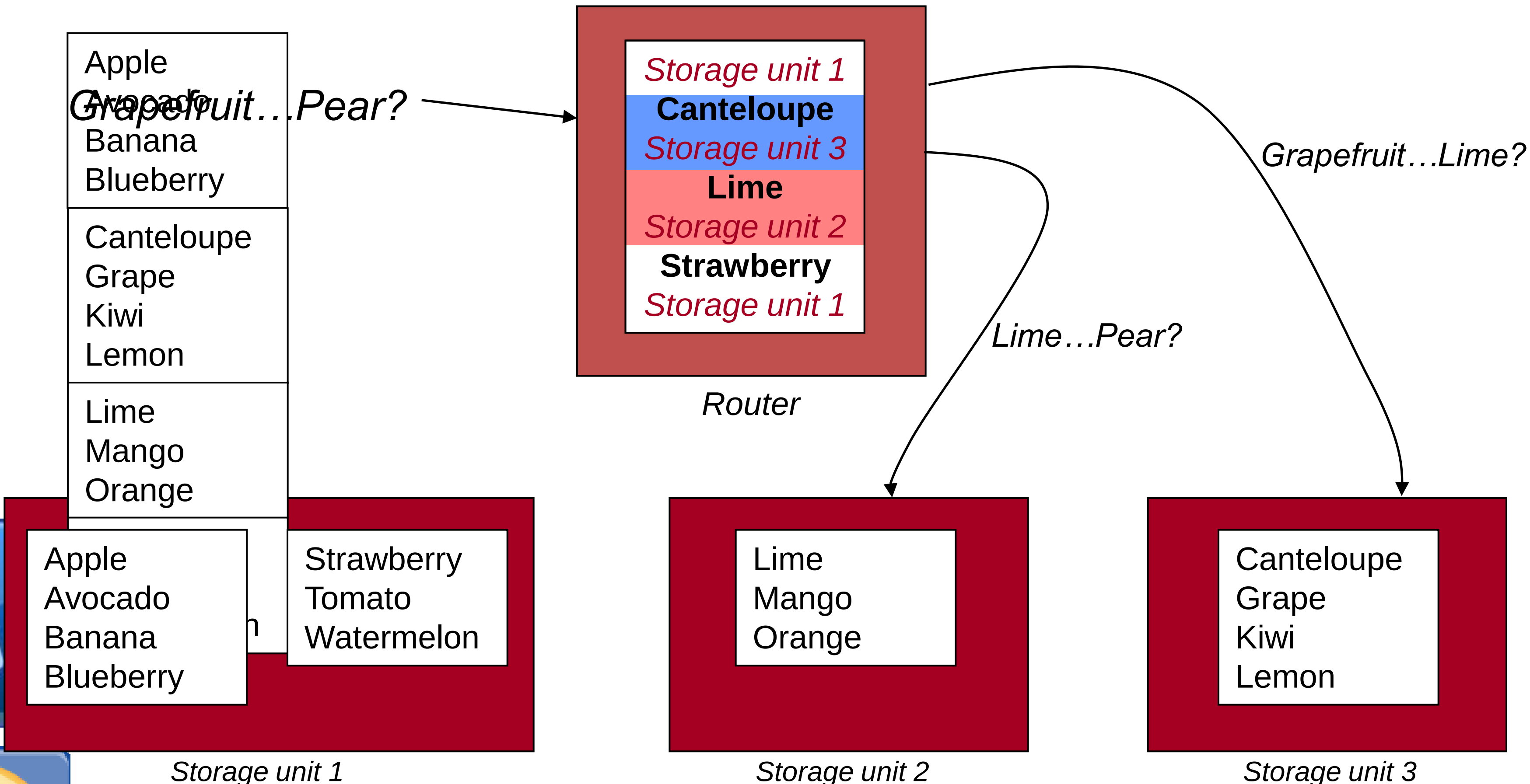
Tablets—Ordered Table

	Name	*Description*	*Price*
A	Apple	Apple is wisdom	$1
	Avocado	But at what price?	$3
	Banana	The perfect fruit	$2
	Grape	Grapes are good to eat	$12
H	Kiwi	New Zealand	$8
	Lemon	How much did you pay for this lemon?	$1
	Lime	Limes are green	$9
	Orange	Arrgh! Don’t get scurvy!	$2
Q	Strawberry	Strawberry shortcake	$900
	Tomato	Is this a vegetable?	$14

Z

Range Queries in YDOT

Clustered, ordered retrieval of records

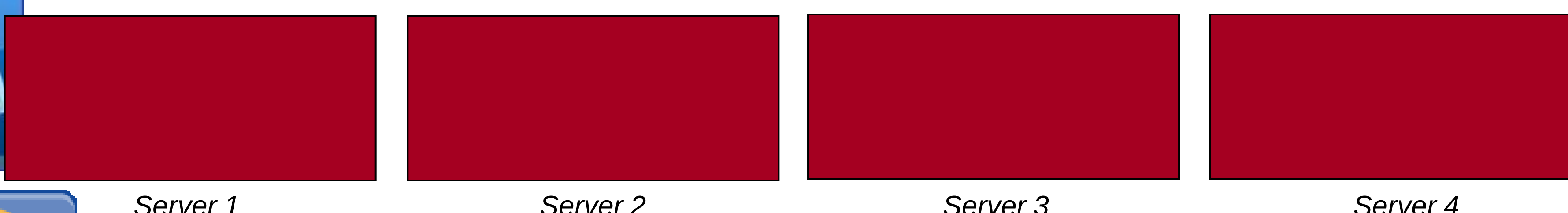


ELASTICITY, OPERABILITY, HORIZONTAL SCALING

Distribution

Data shuffling for load balancing

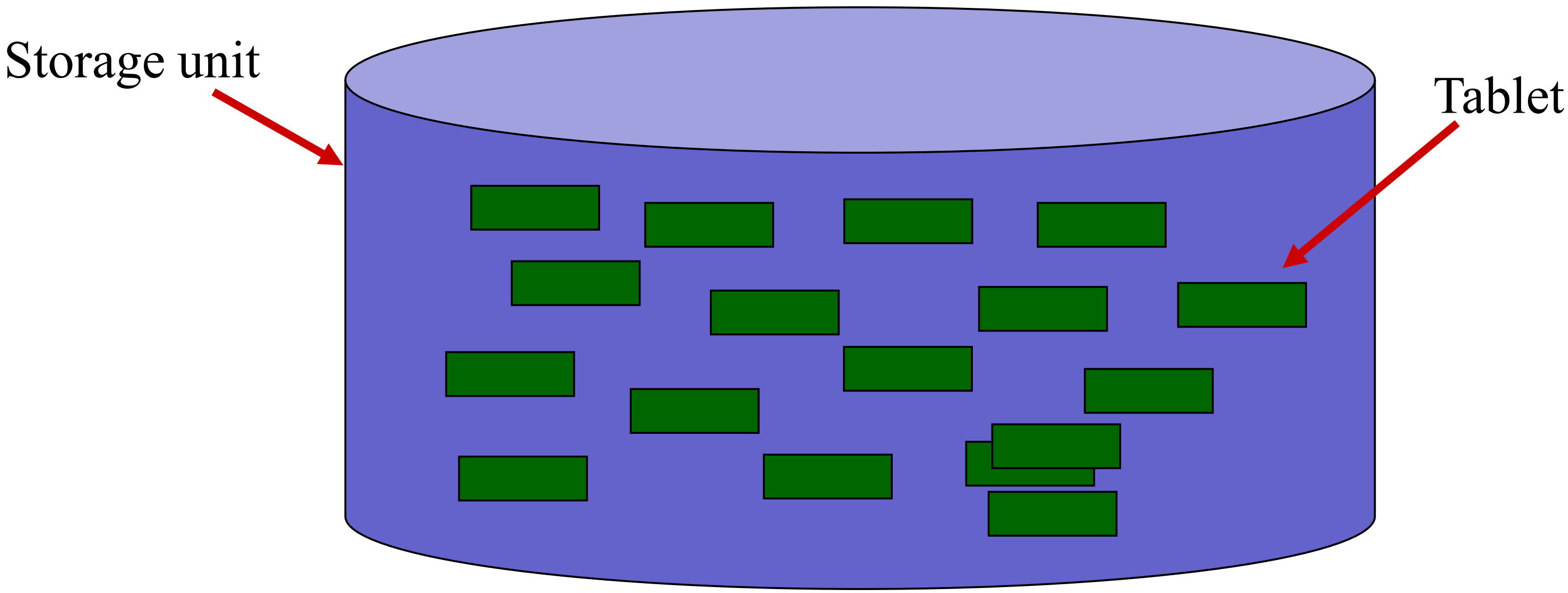
6/1/07	[illegible]	[illegible]	$70
6/1/07	256623	Car	$1123
6/2/07	636353	Bike	$86
6/5/07	662113	Chair	$10
6/7/07	121113	Lamp	$19
6/9/07	887734	Bike	$56
6/11/07	252111	Scooter	$18
6/11/07	116458	Hammer	$8000



Tablet Splitting and Balancing

Each storage unit has many tablets (horizontal partitions of the table)

Storage unit may become a hotspot



Overfull tablets split

Tablets may grow over time

Shed load by moving tablets to other servers

ASYNCHRONOUS REPLICATION AND CONSISTENCY

Asynchronous Replication

Consistency: Social Alice

West | **East** | **Record Timeline**

User	Status
Alice	Busy

Network disruption:
Alice redirected to East

East:

User	Status
Alice	Free

West:

User	Status
Alice	Busy

busy →

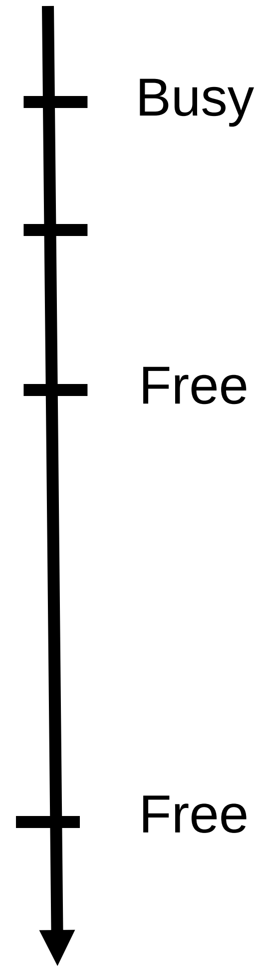
← **free**

West:

User	Status
Alice	???

East:

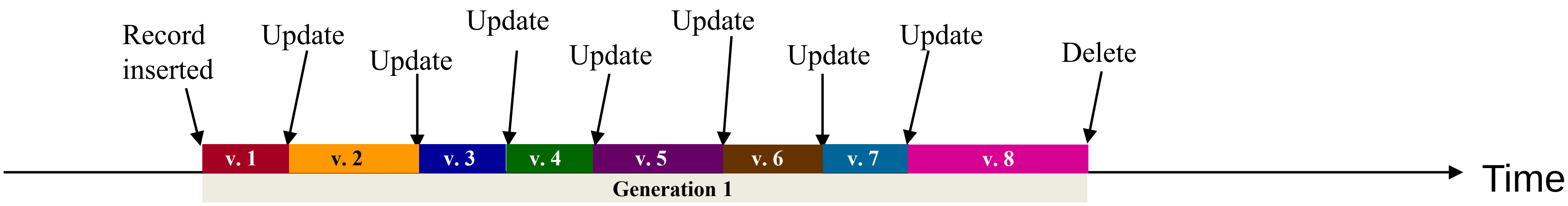
User	Status
Alice	???



PNUTS Consistency Model

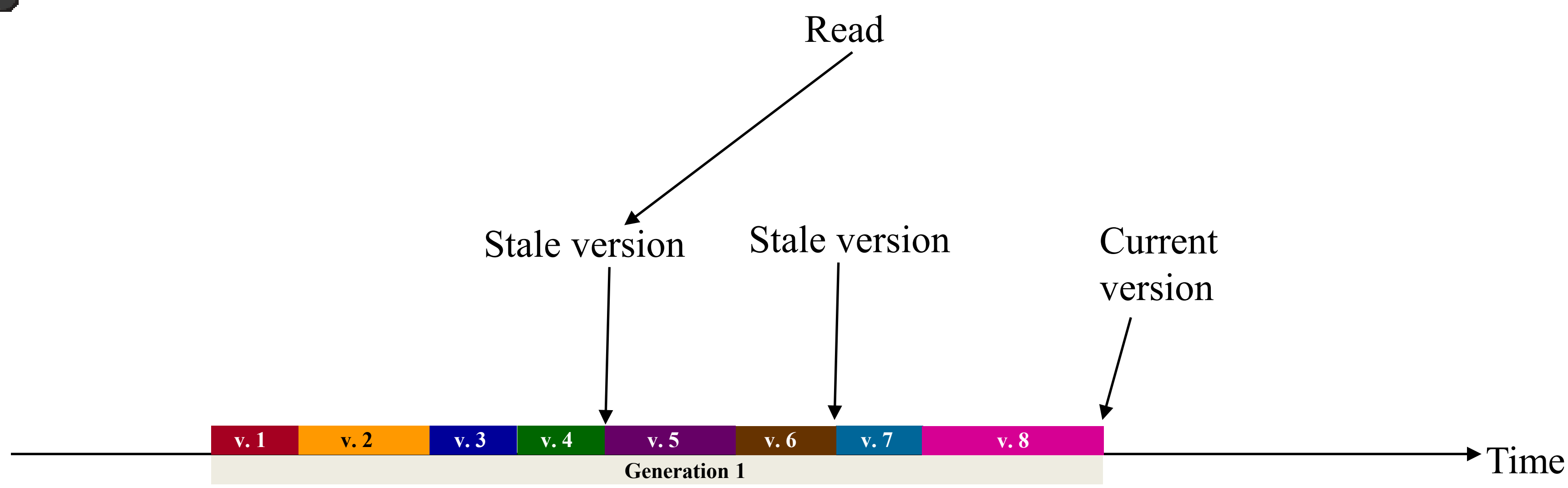
Goal: Make it easier for applications to reason about updates and cope with asynchrony

What happens to a record with primary key "Alice"?



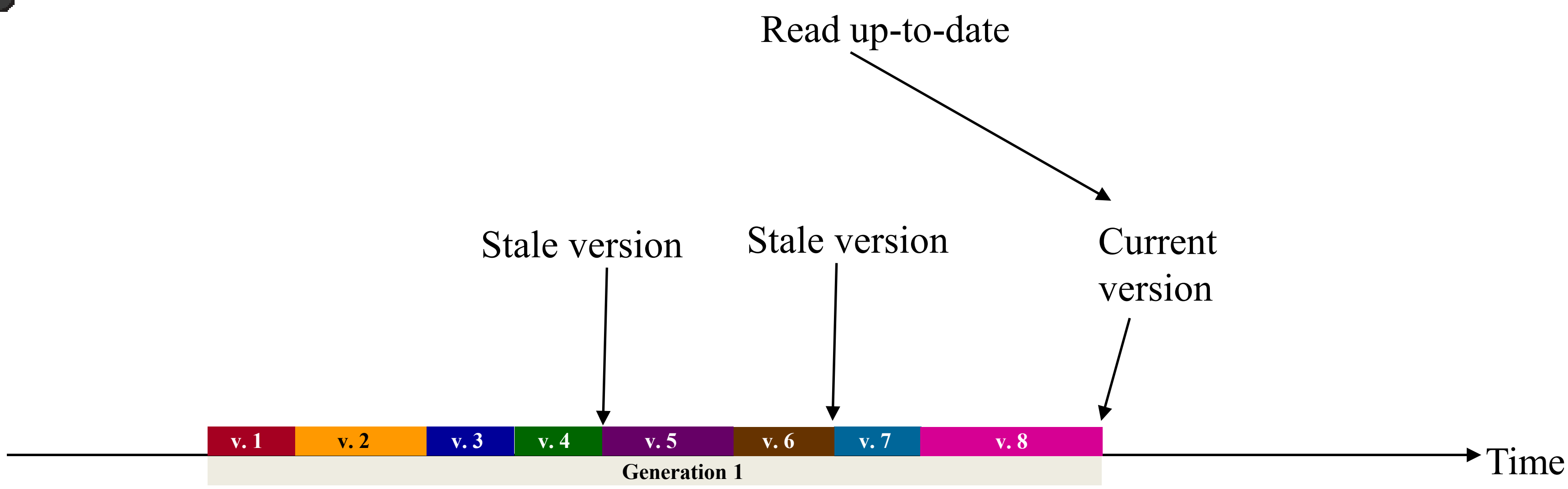
As the record is updated, copies may get out of sync.

PNUTS Consistency Model



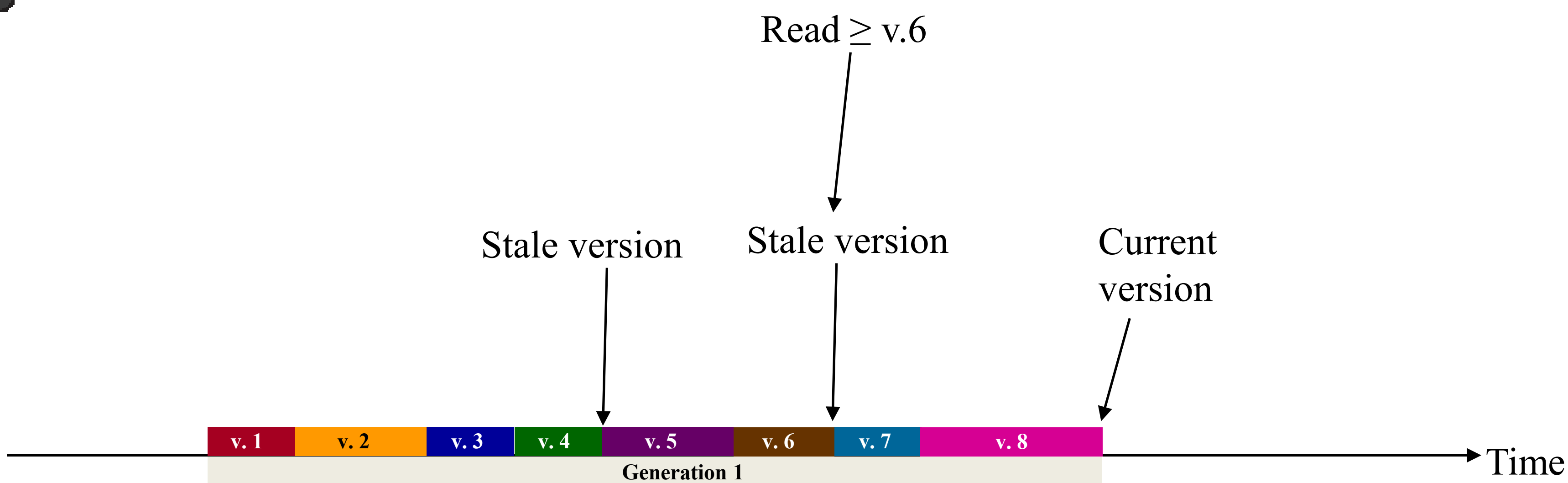
In general, reads are served using a local copy

PNUTS Consistency Model



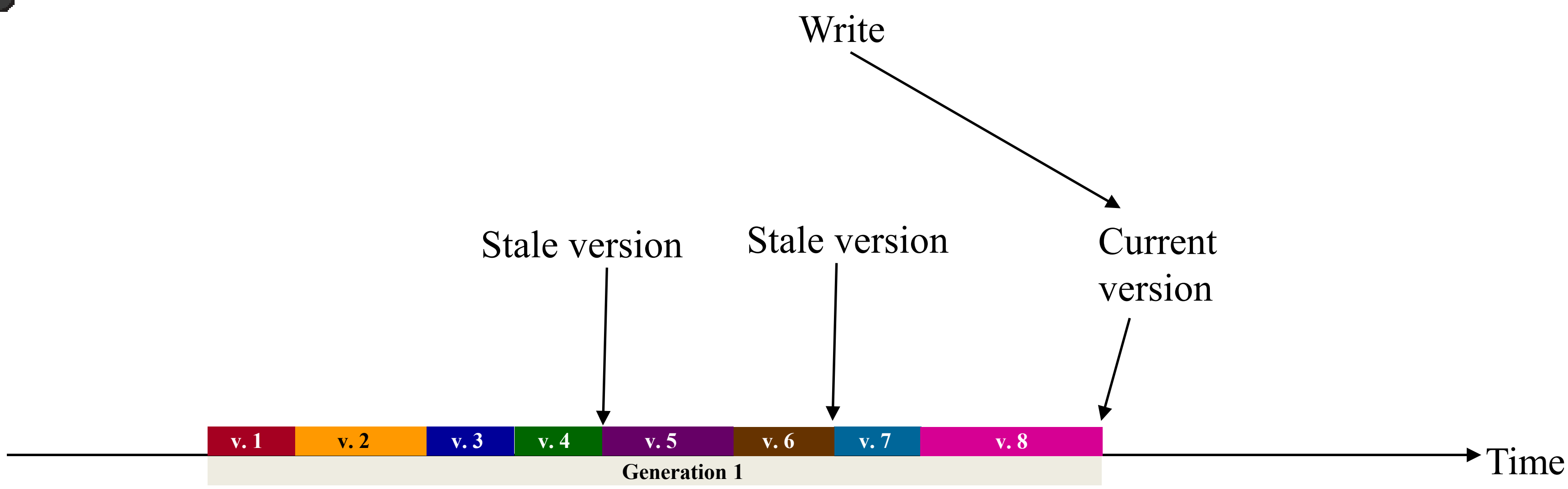
But application can request and get current version

PNUTS Consistency Model



Or variations such as “read forward”—while copies may lag the master record, *every* copy goes through the same sequence of changes

PNUTS Consistency Model

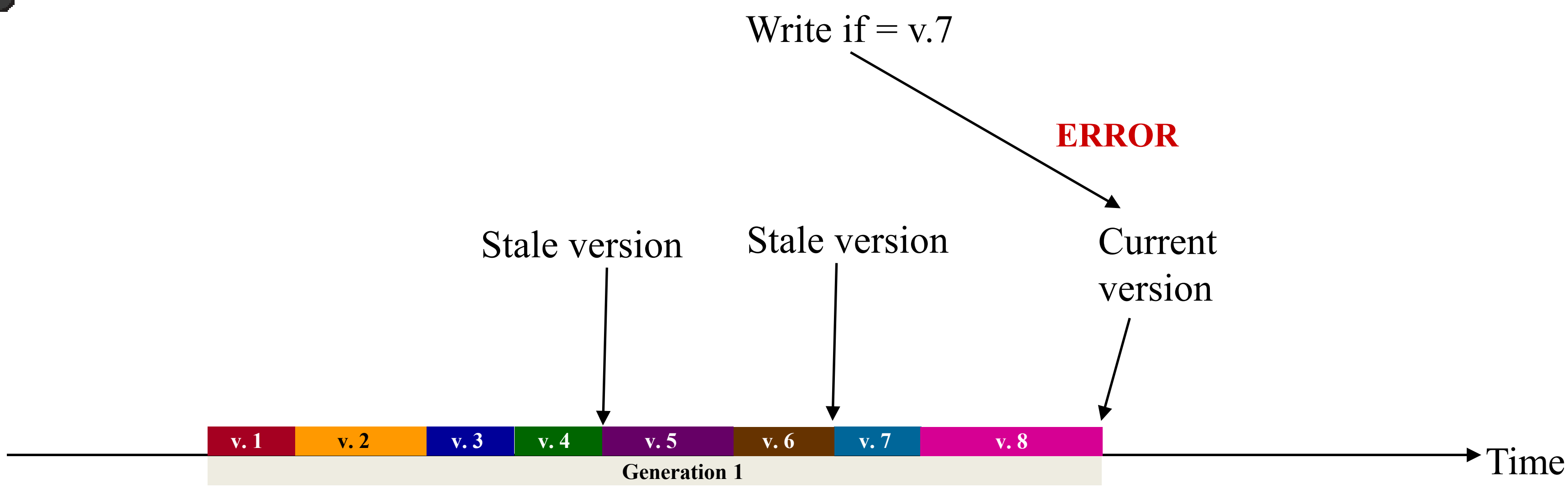


Achieved via per-record primary copy protocol

(To maximize availability, record masterships automatically transferred if site fails)

Can be selectively weakened to eventual consistency
(local writes that are reconciled using version vectors)

PNUTS Consistency Model



Test-and-set writes facilitate per-record transactions

Consistency Techniques

Per-record mastering

- Each record is assigned a “master region”
  - May differ between records
- Updates to the record forwarded to the master region
- Ensures consistent ordering of updates

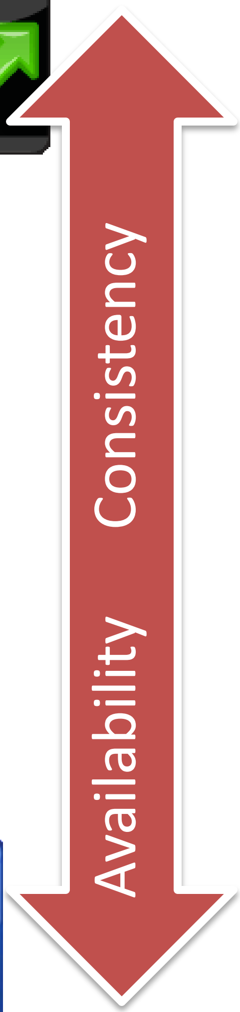
Tablet-level mastering

- Each tablet is assigned a “master region”
- Inserts and deletes of records forwarded to the master region
- Master region decides tablet splits

These details are hidden from the application

- Except for the latency impact!

Consistency Levels



Primary Key Constraint + Record Timeline

- Each tablet is assigned a “master region”
- Inserts of records forwarded to the master region
- Inserts and updates could fail during outages*

Record Timeline Consistency

- Each record is assigned a “master region”
- Updates to the record forwarded to the master region
- Inserts succeed, but updates could fail during outages*

Eventual Consistency

- Low latency updates and inserts done locally
- Per field timestamp used to merge updates

★ In case of SU or data center failure. We have failover tools!

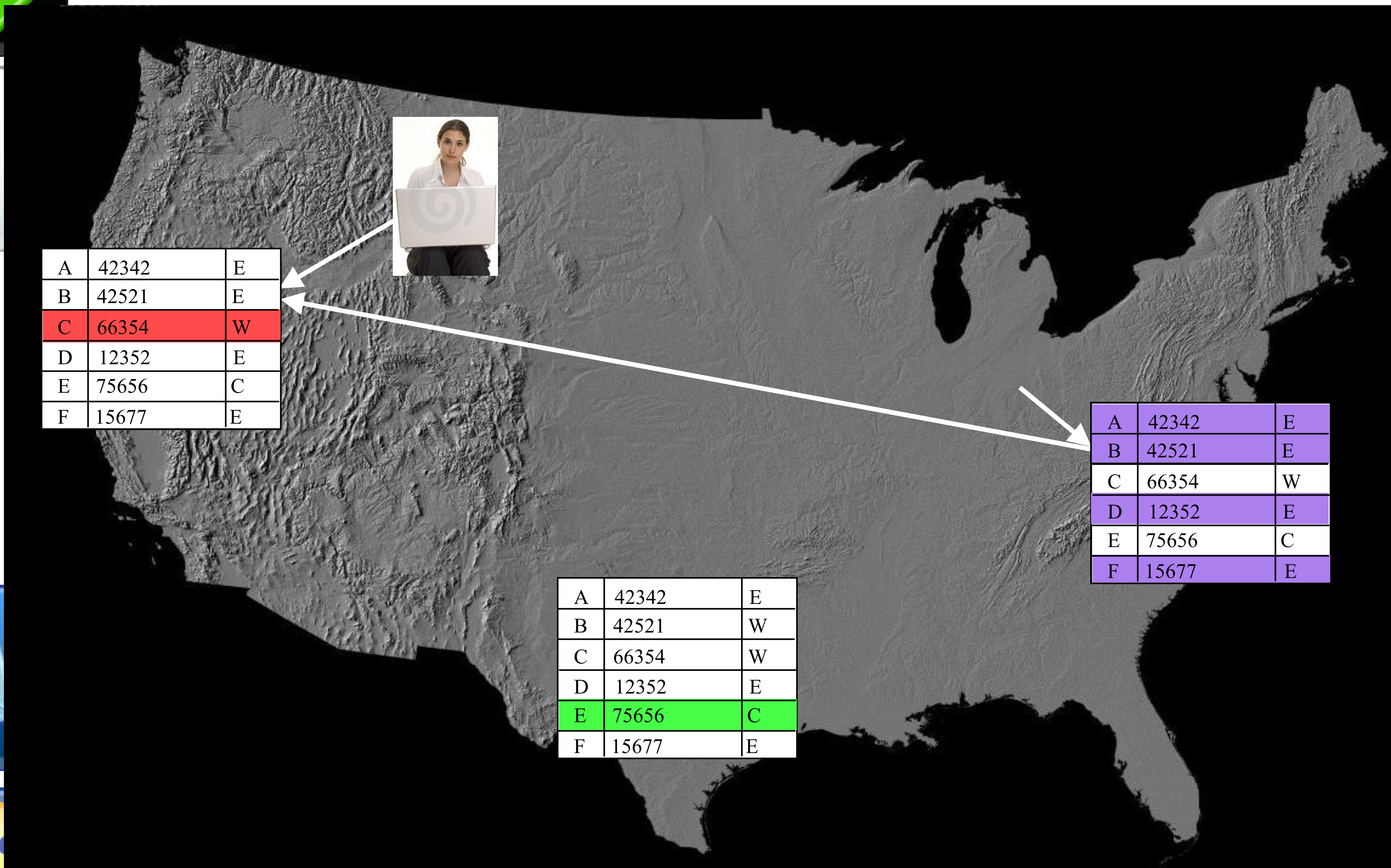
★ Reads always will be sent to another region

Generalizing Record Timelines to Partition Timelines

Record ➡ Partitition of records with same key

- Tablet splits must respect partition boundaries
- Intra-partition ACID transactions can be done easily now
  - Single machine transactions!
  - With composite keys, this captures Azure and Google AE models
- Each partition is assigned a “master region”
  - May differ between partitions
- Updates to the partition forwarded to the master region
- Ensures consistent ordering of updates across nodes

Record Master

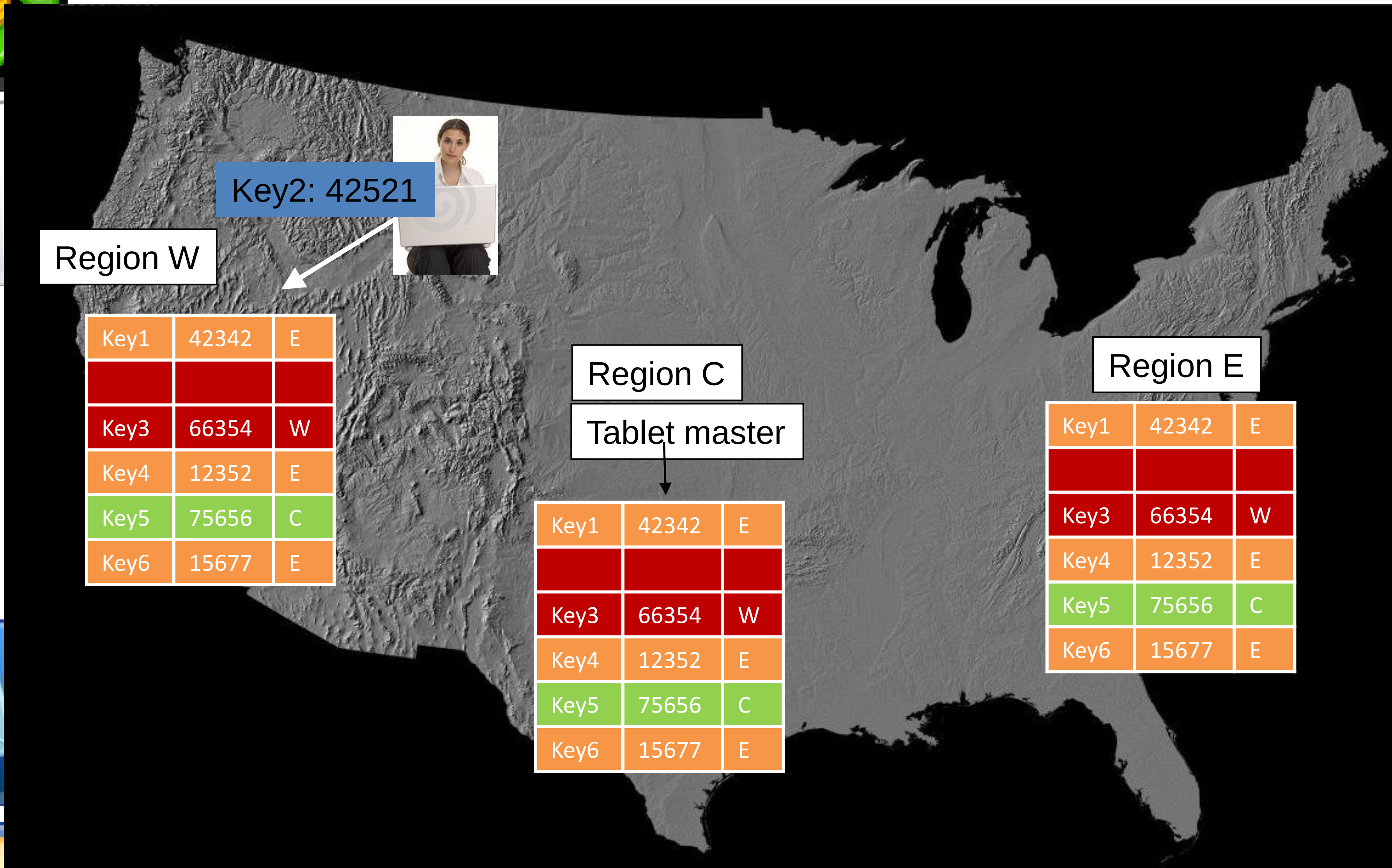


A	42342	E
B	42521	E
C	66354	W
D	12352	E
E	75656	C
F	15677	E

A	42342	E
B	42521	E
C	66354	W
D	12352	E
E	75656	C
F	15677	E

A	42342	E
B	42521	W
C	66354	W
D	12352	E
E	75656	C
F	15677	E

Tablet Master



Region W:

Key1	42342	E
Key3	66354	W
Key4	12352	E
Key5	75656	C
Key6	15677	E

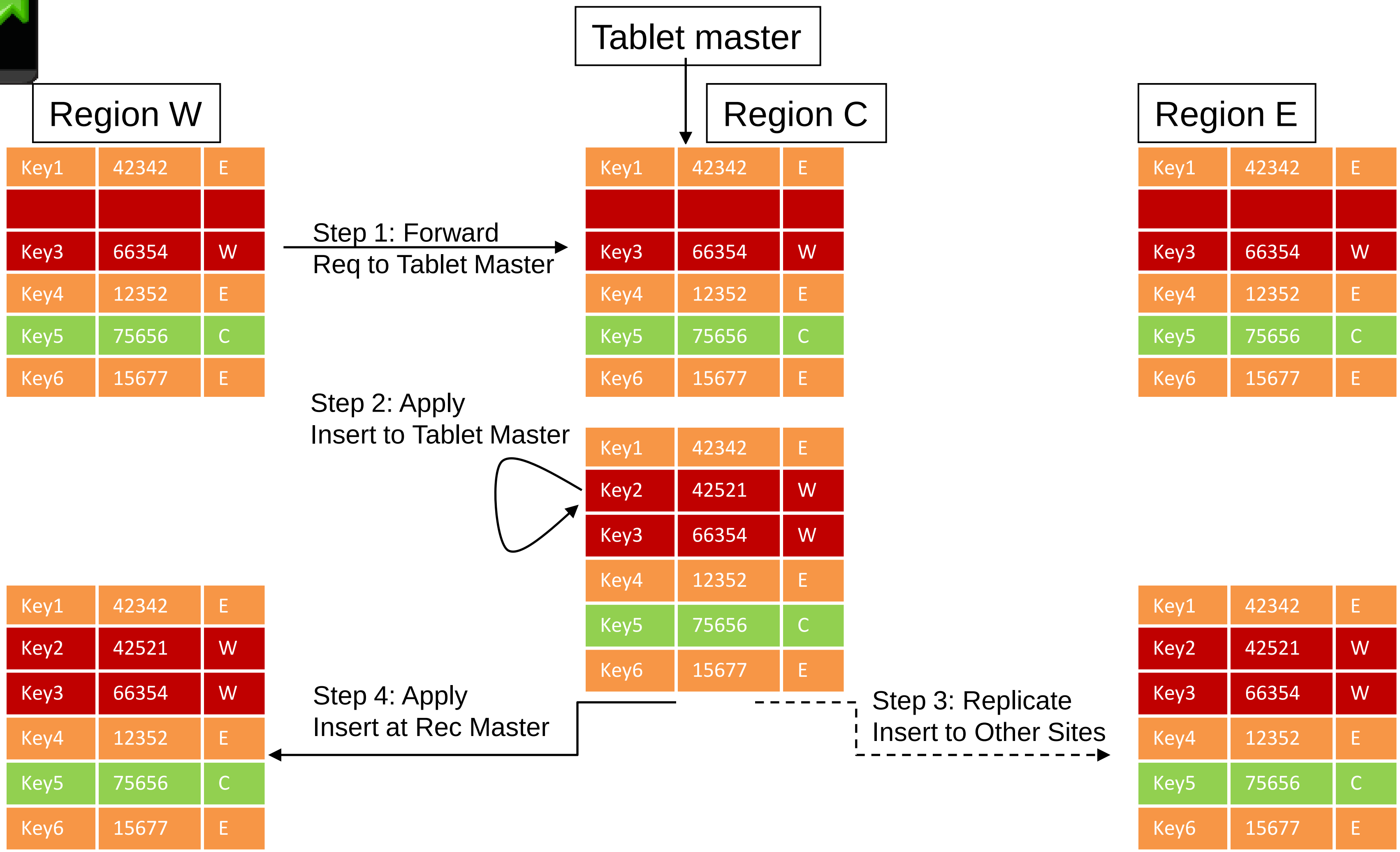
Region C (Tablet master):

Key1	42342	E
Key3	66354	W
Key4	12352	E
Key5	75656	C
Key6	15677	E

Region E:

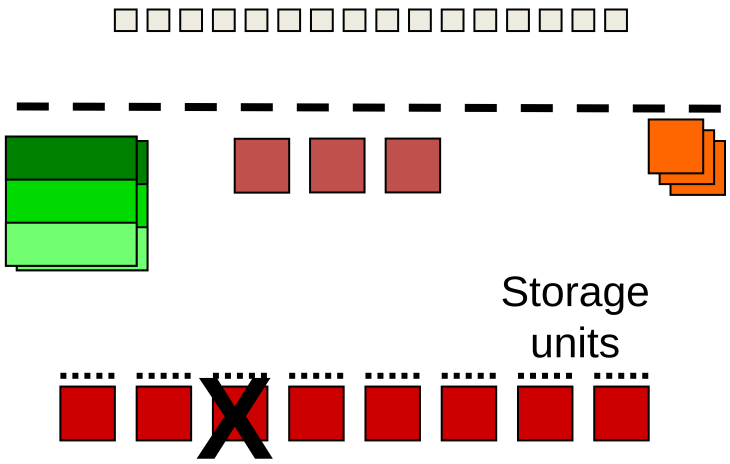
Key1	42342	E
Key3	66354	W
Key4	12352	E
Key5	75656	C
Key6	15677	E

Tablet Mastership



AVAILABILITY

Possible Failure Modes



Failure type

Storage unit

Consistency impact

None

Availability impact

Degraded service (forwards) for some data.
Updates and inserts fail for some records

Resolution

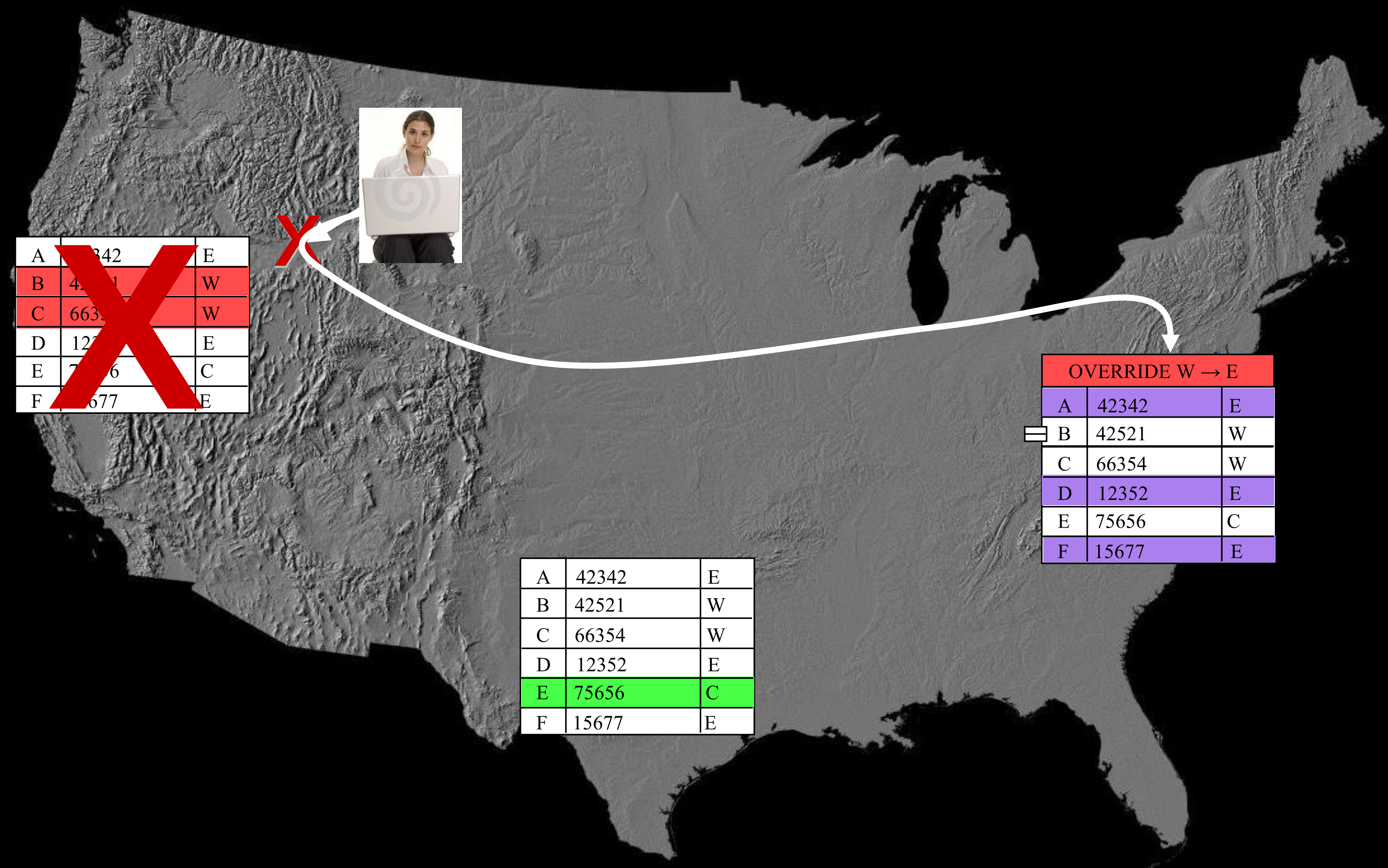
If data not lost: Reboot machine

If data lost: Copy lost tablets from a remote replica

Time to resolve

If data lost, hours or less (depending on tablet size and colo location). If no data lost, minutes.

Coping With Failures



A	342	E
B	42 1	W
C	663	W
D	12	E
E	7 6	C
F	677	E

OVERRIDE W → E		
A	42342	E
B	42521	W
C	66354	W
D	12352	E
E	75656	C
F	15677	E

A	42342	E
B	42521	W
C	66354	W
D	12352	E
E	75656	C
F	15677	E

Possible Failure Modes

Failure type

Router

Consistency impact

None

Availability impact

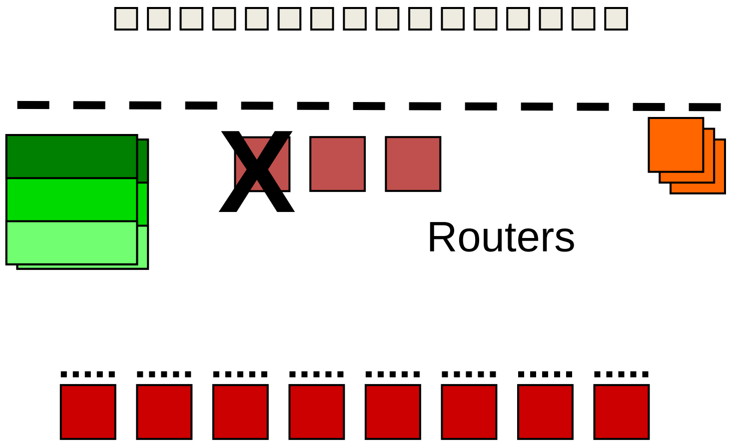
None

Resolution

Boot router

Time to resolve

Minutes



Possible Failure Modes

Failure

Tablet controller

Consistency impact

None

Availability impact

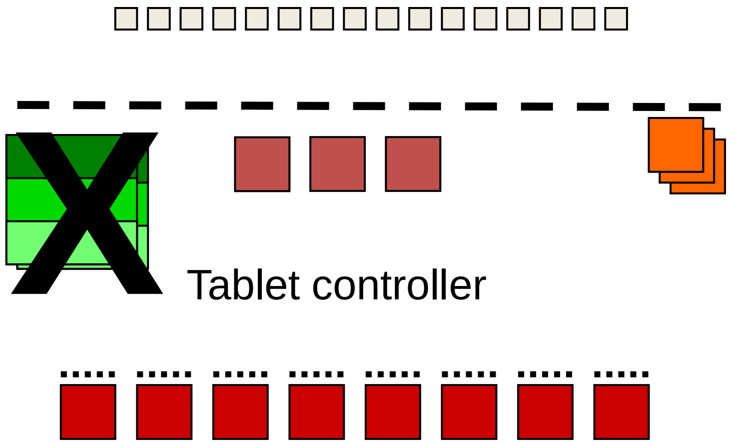
Some actions (e.g., tablet copy) will be blocked

Resolution

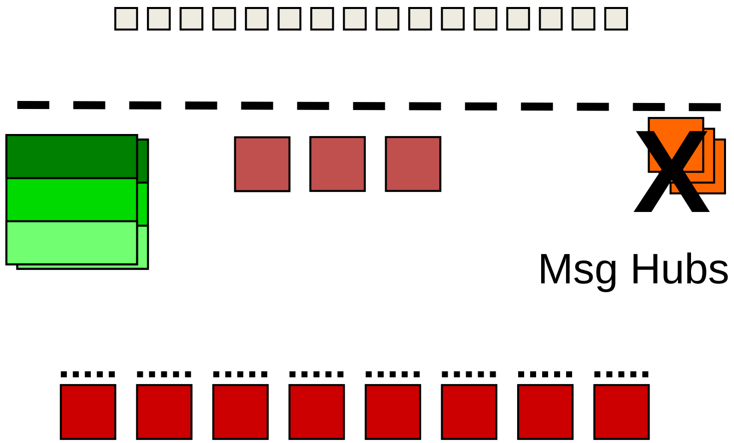
Start secondary controller

Time to resolve

Minutes



Possible Failure Modes



Failure

One msg hub node

Consistency impact

None

Availability impact

Writes fail for some records until a new secondary node takes over

Resolution

Create new primary or secondary for lost topics

Time to resolve

Minutes

Possible Failure Modes

Failure

Colo power outage or partition

Consistency impact

Option to allow "relaxed consistency" to improve availability

Availability impact

Some inserts, updates and deletes cannot succeed

Some critical reads fail

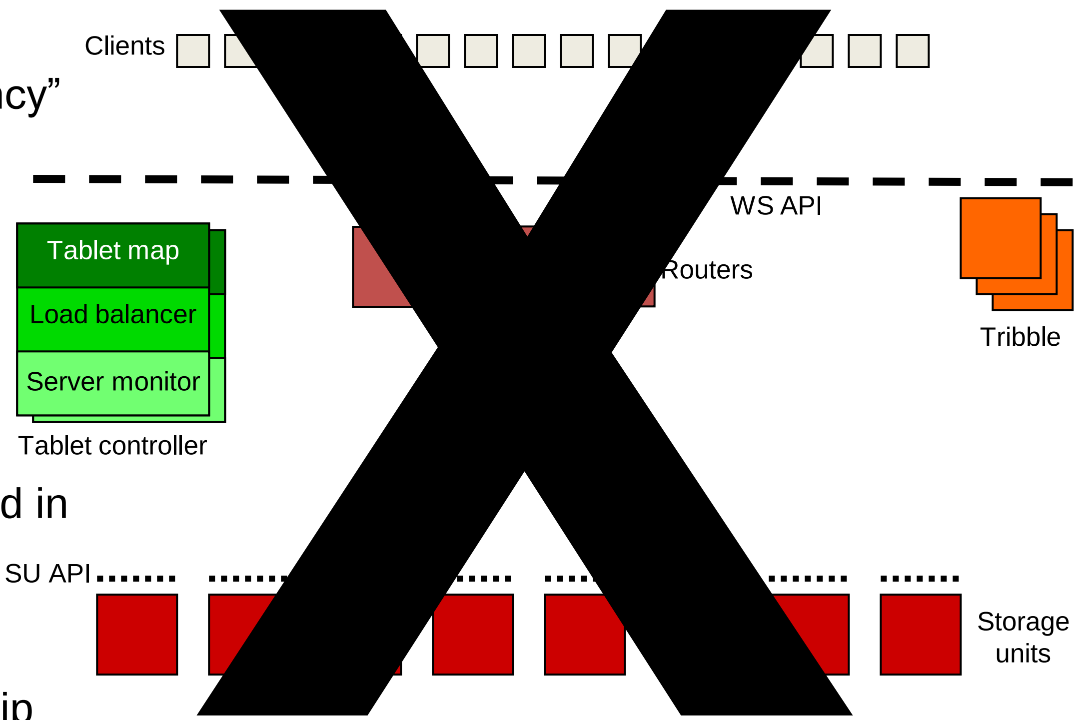
Option to allow updates to proceed in "relaxed consistency mode"

Resolution

Major overrides to force mastership transfer; discard conflicting updates

Time to resolve

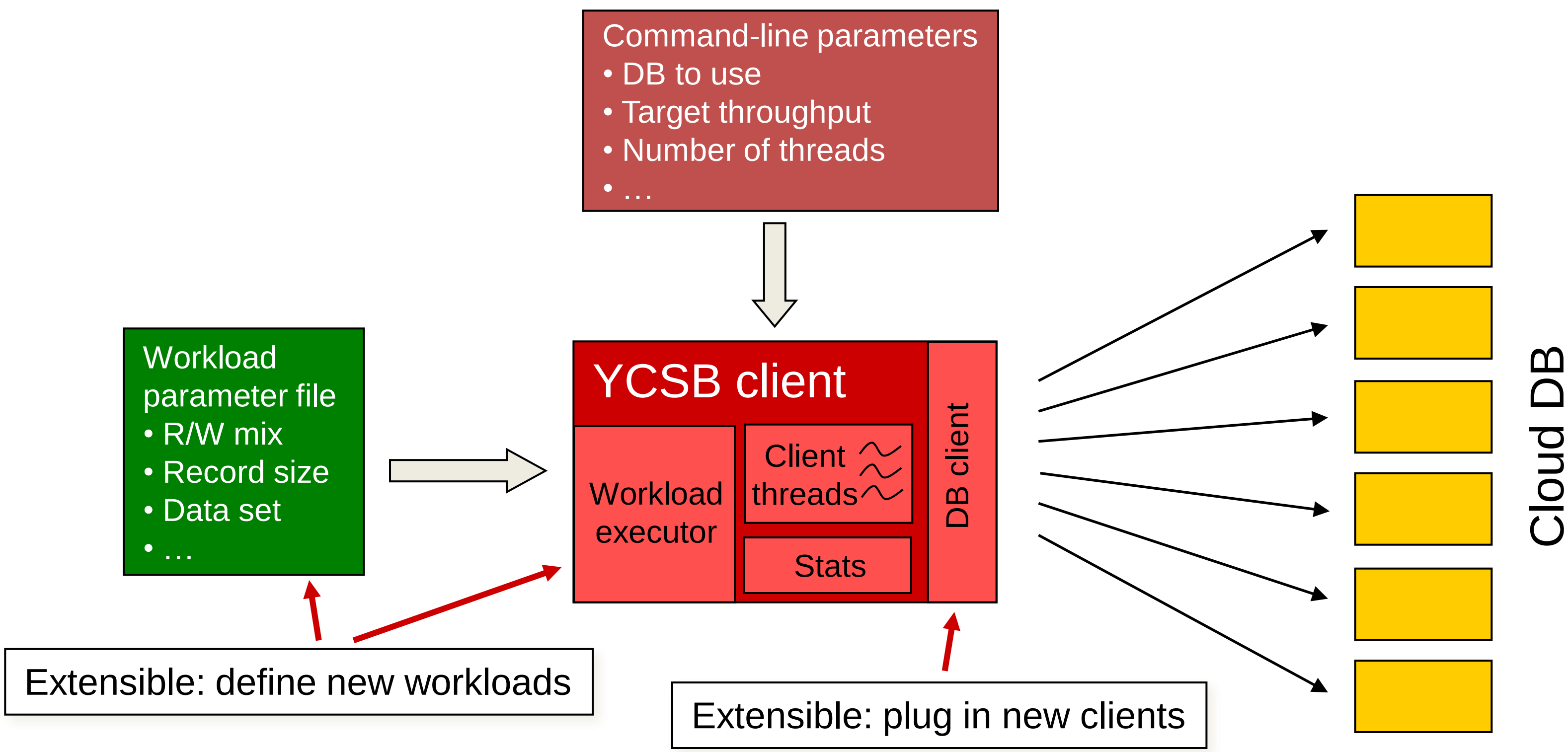
Hours



YCS Benchmark Tool

Java application

- Many systems have Java APIs
- Other systems via HTTP/REST, JNI or some other solution



Walnut

How should next-gen Yahoo! cloud be architected?

Further PNutty Reading

Efficient Bulk Insertion into a Distributed Ordered Table (SIGMOD 2008)
Adam Silberstein, Brian Cooper, Utkarsh Srivastava, Erik Vee,
Ramana Yerneni, Raghu Ramakrishnan

PNUTS: Yahoo!'s Hosted Data Serving Platform (VLDB 2008)
Brian Cooper, Raghu Ramakrishnan, Utkarsh Srivastava,
Adam Silberstein, Phil Bohannon, Hans-Arno Jacobsen,
Nick Puz, Daniel Weaver, Ramana Yerneni

Asynchronous View Maintenance for VLSD Databases (SIGMOD 2009)
Parag Agrawal, Adam Silberstein, Brian F. Cooper, Utkarsh Srivastava and
Raghu Ramakrishnan

Cloud Storage Design in a PNUTShell
Brian F. Cooper, Raghu Ramakrishnan, and Utkarsh Srivastava
Beautiful Data, O'Reilly Media, 2009

Adaptively Parallelizing Distributed Range Queries (VLDB 2009)
Ymir Vigfusson, Adam Silberstein, Brian Cooper, Rodrigo Fonseca