CRDTs

Data Types for EC Systems

Problem?

Eventual Consistency

Eventual consistency is a consistency model used in distributed computing that informally guarantees that, if no new updates are made to a given data item, eventually all accesses to that item will return the last updated value.

--Wikipedia

Trade Off

Scale Up

$$$Big Iron
(still fails)

Scale Out

Commodity Servers
CDNs, App servers
Expertise

Fault Tolerance

Low Latency

Low Latency

Amazon found every 100ms of latency cost them 1% in sales.

Low Latency

Low Latency

Google found an extra 0.5 seconds in search page generation time dropped traffic by 20%.

CAP

C

A

EC

Causal
RYOW
Session
Monotonic Read/Write

Pick Your Own

http://basho.com/understanding-riaks-configurable-behaviors-part-1/

http://www.bailis.org/blog/when-is-acid-acid-rarely/

Who Pays?

Developers

But how?

Google F1

“We have a lot of experience with eventual consistency systems at Google.”

“We find developers spend a significant fraction of their time building extremely complex and error-prone mechanisms to cope with eventual consistency”

Riak Overview

Riak Overview

Erlang implementation of Dynamo

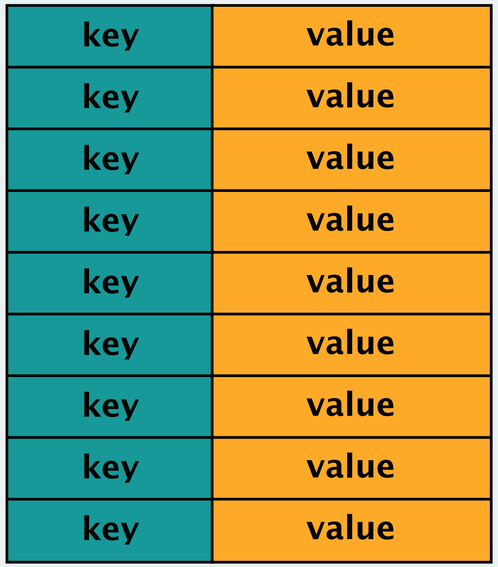

riak

{“key”:
“value”}

{“key”:
“value”}

Keys are namespaced

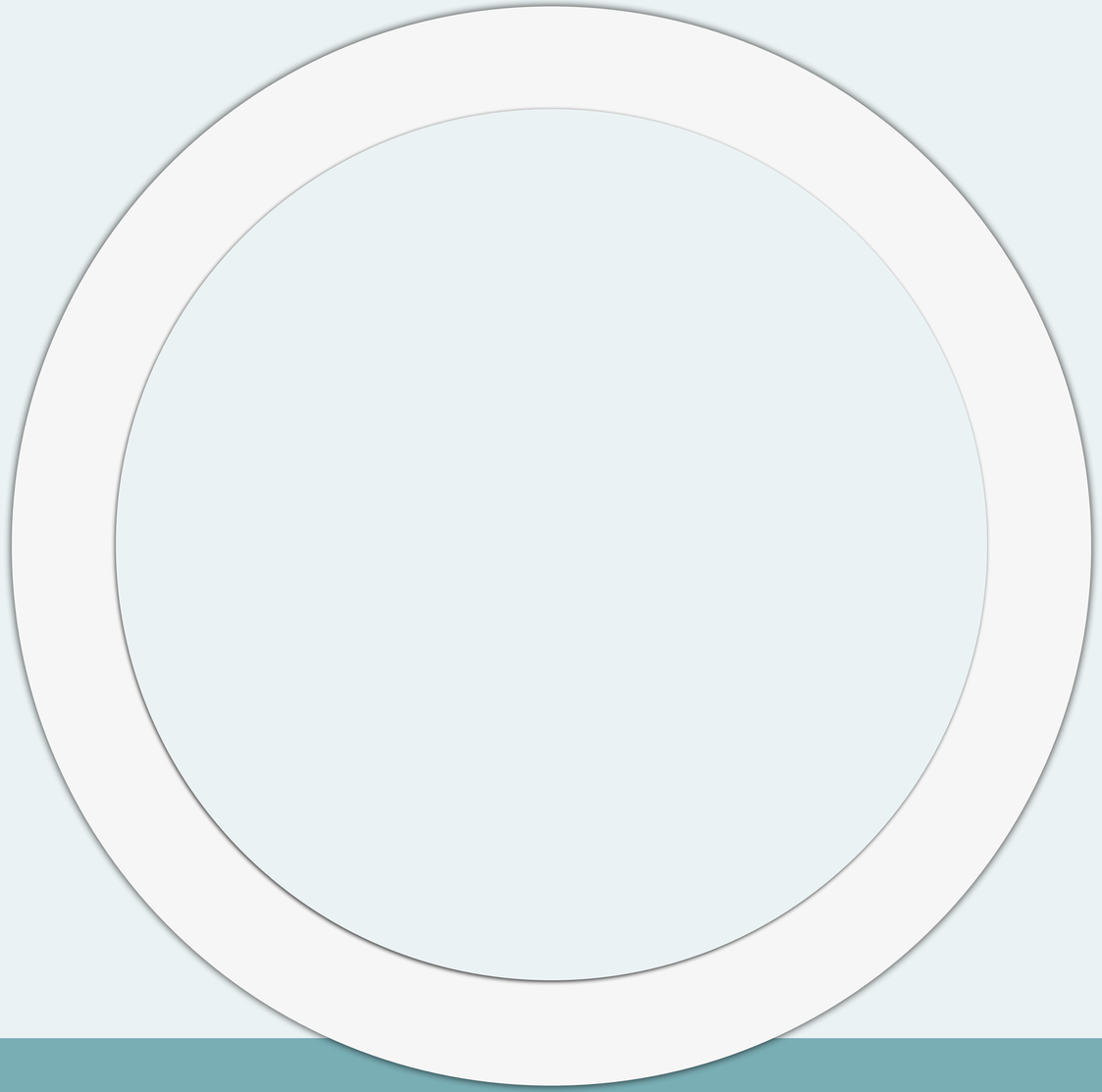
into **Buckets**



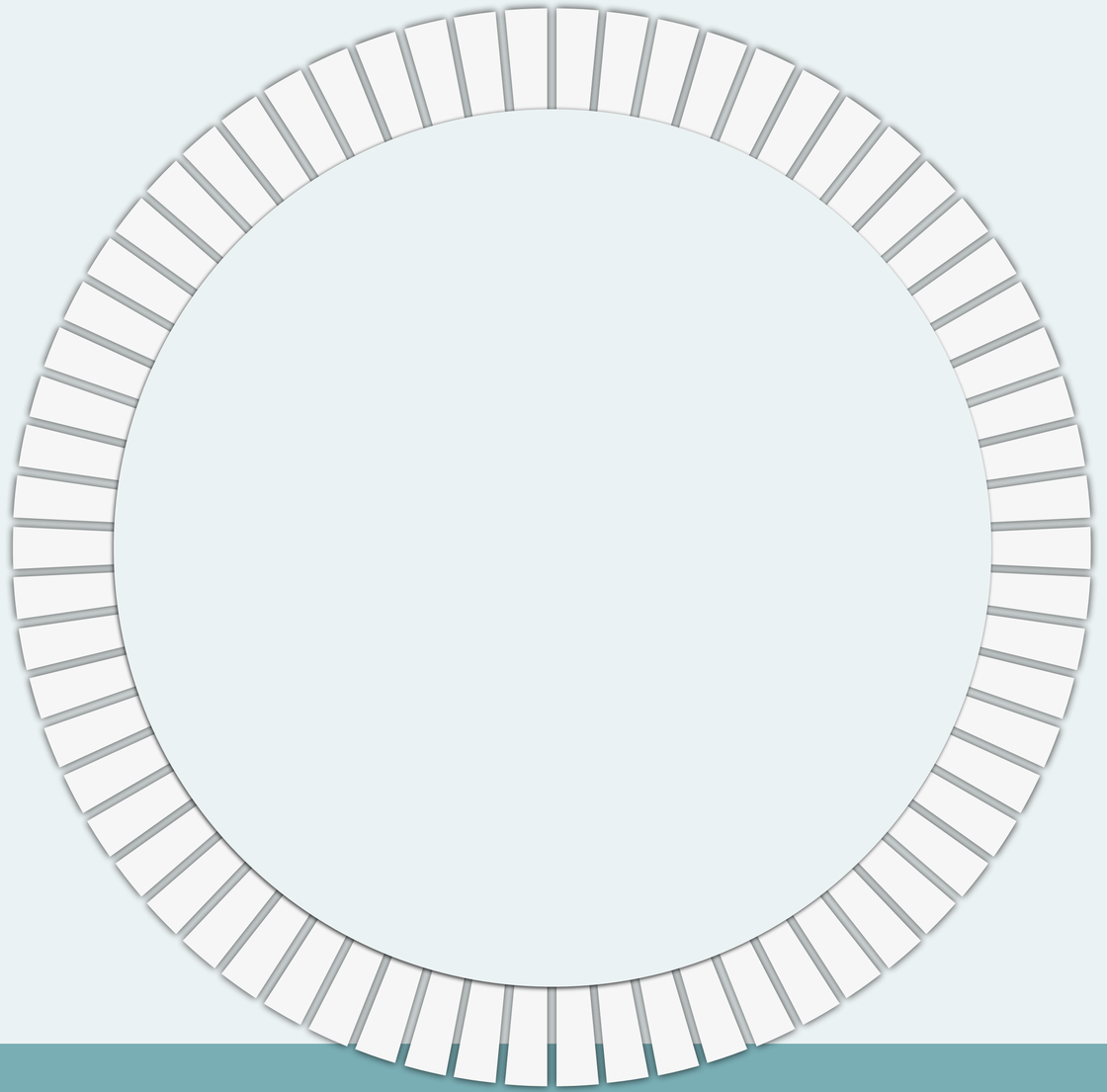
riak

Riak Overview

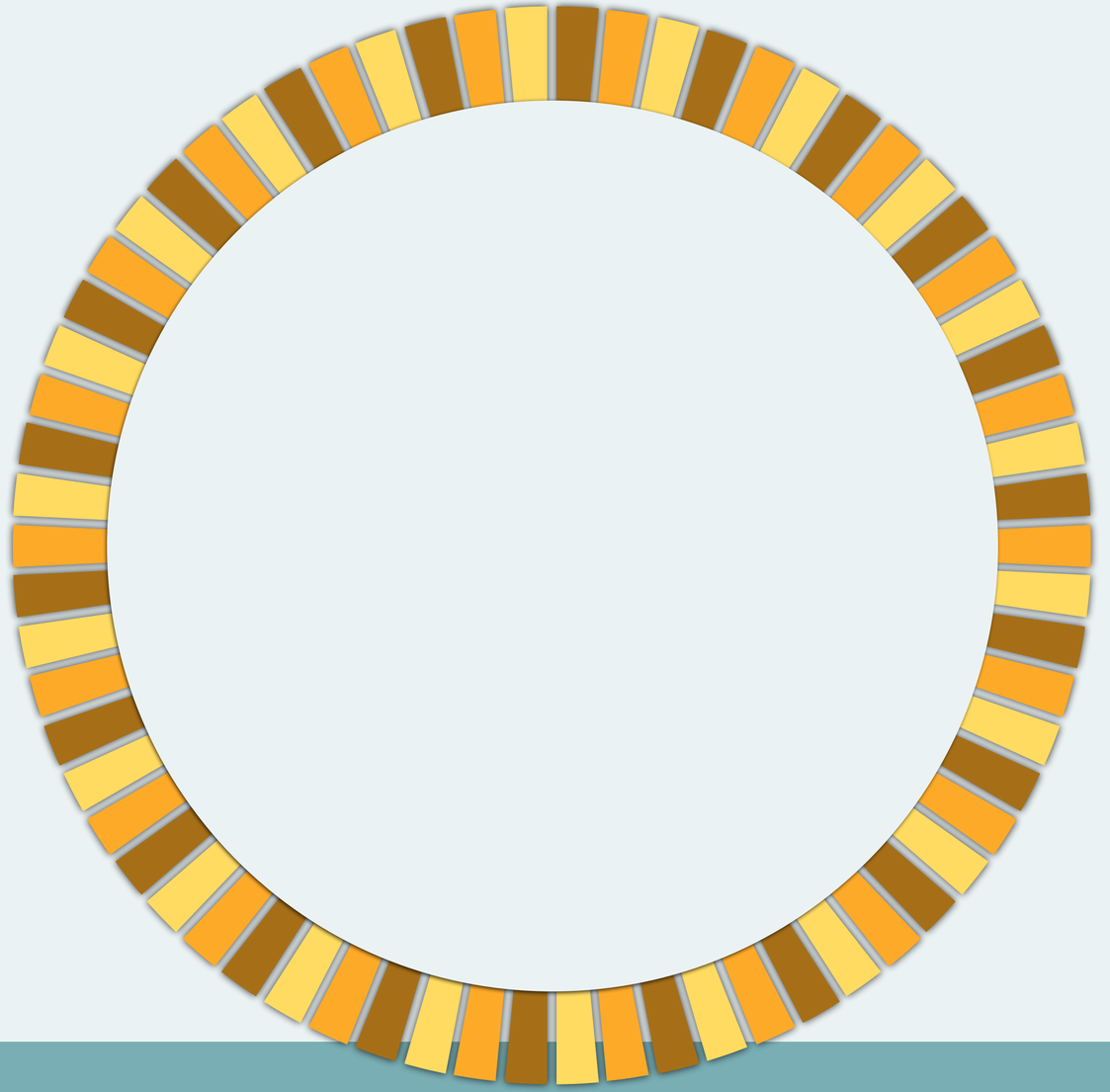
Consistent Hashing



riak



riak



riak

Riak Overview

Dynamic Membership

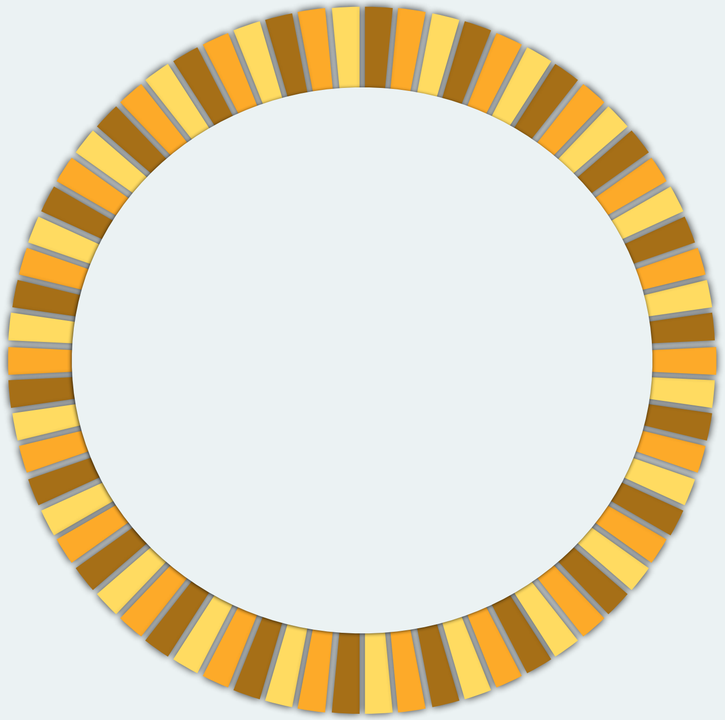
Riak Overview

Replication factor

Replica

Replica

Replica



High Availability

Any non-failing node can respond to any request.

--Gilbert & Lynch

Riak Overview

Two Writes: {Writer, Value, Time}

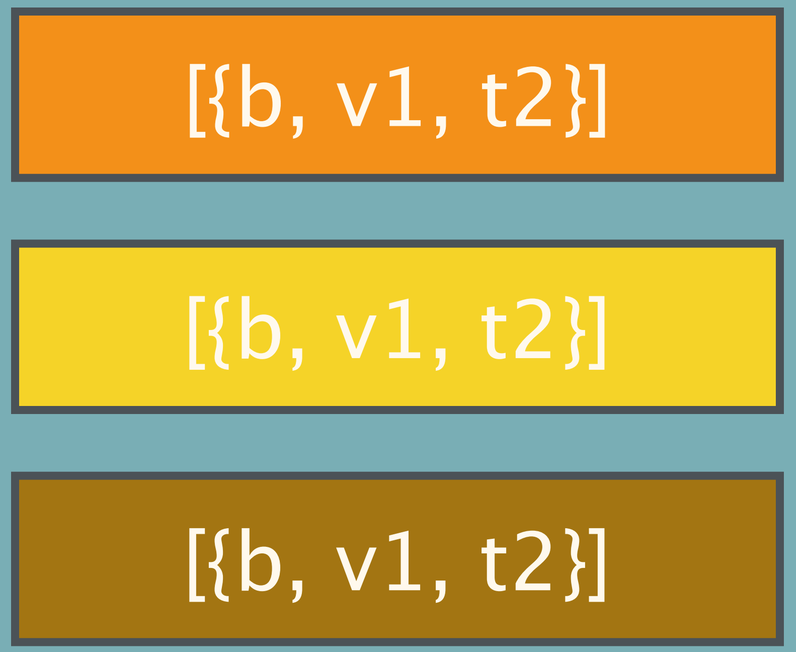

[{a, v1, a1}]
[{b, v2, b1}]
[{a, v1, a1}]

Riak Overview

Last Writer Wins
Allow Mult

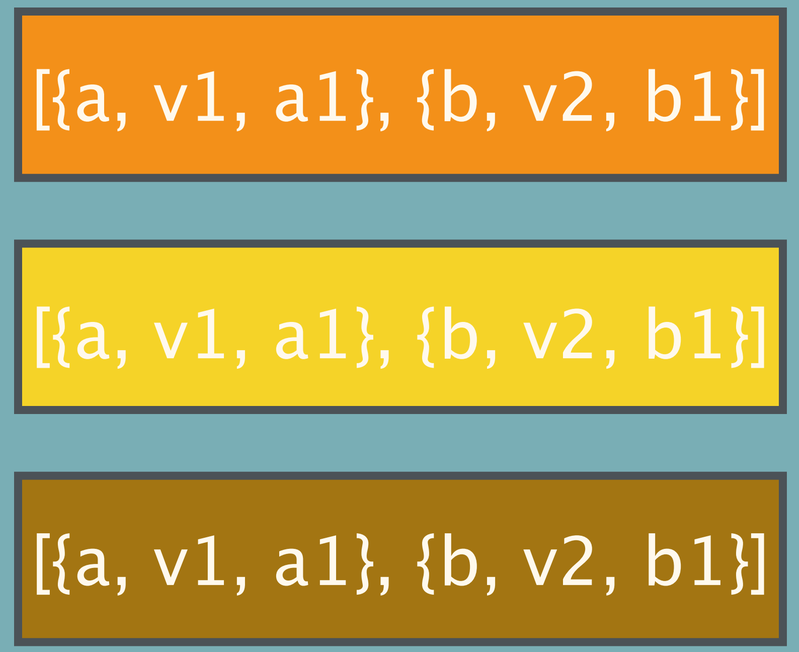
Riak Overview

Last Writer Wins



Riak Overview

Allow Mult



User specified

Merge

Semantic Resolution

DB IS DOWN

Fill in Form P17QR-35
File for input

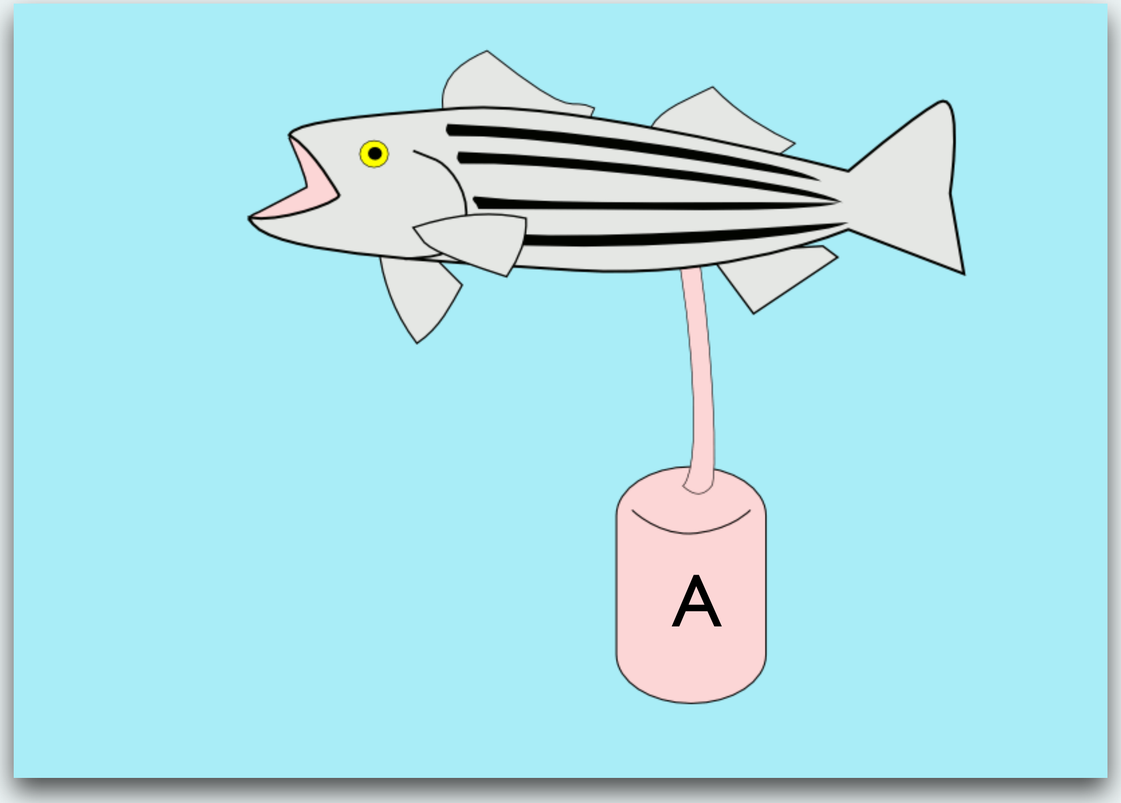
DB IS DOWN

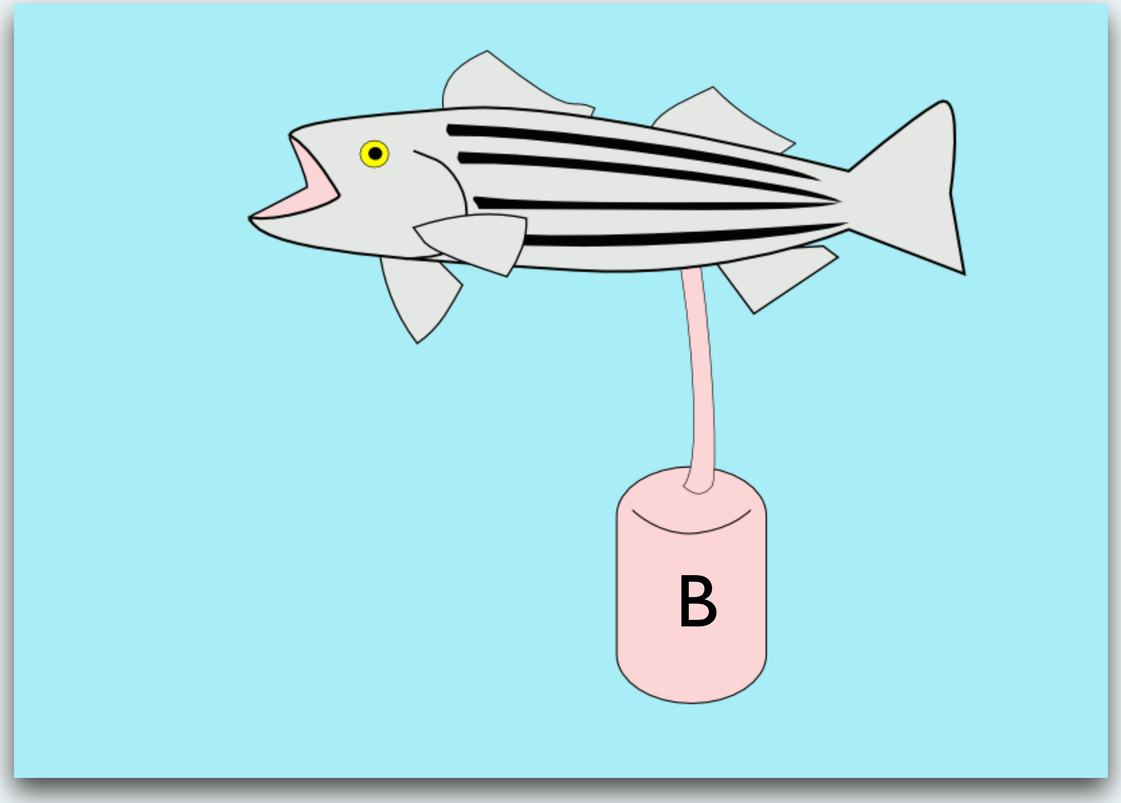
Fill in Form P17QR-35
File for input

OOPS

Dynamo

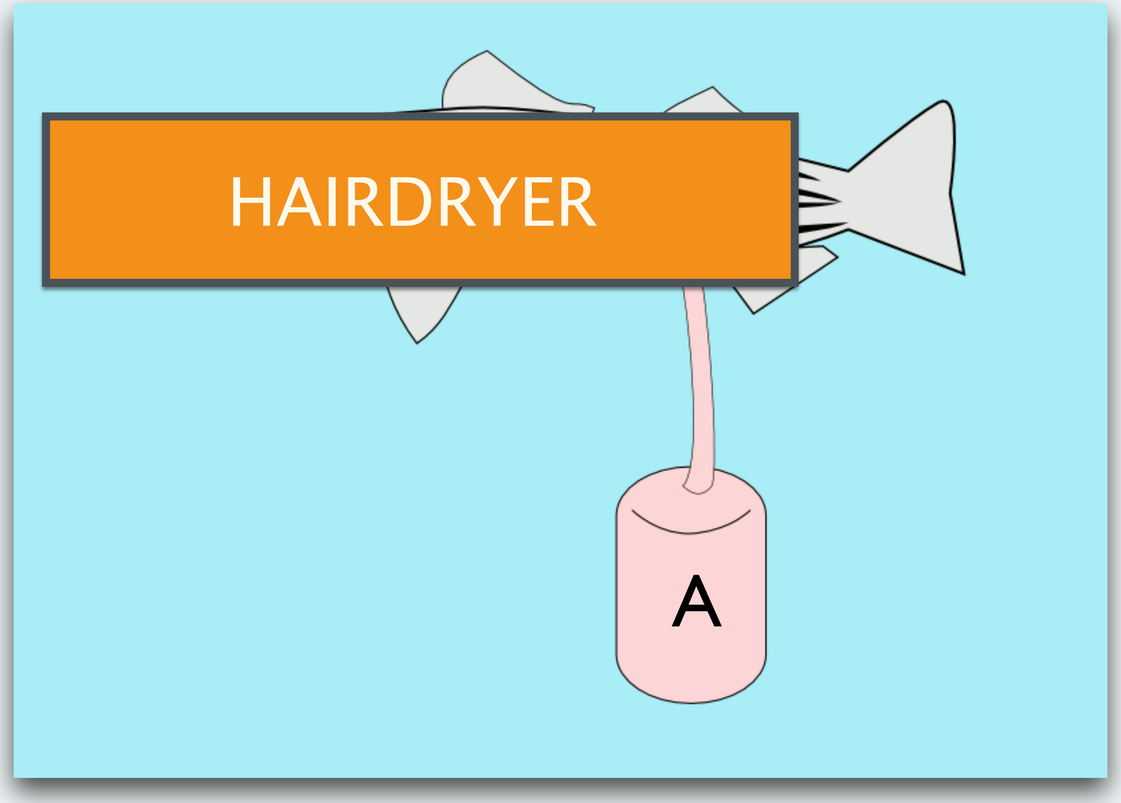
The Shopping Cart

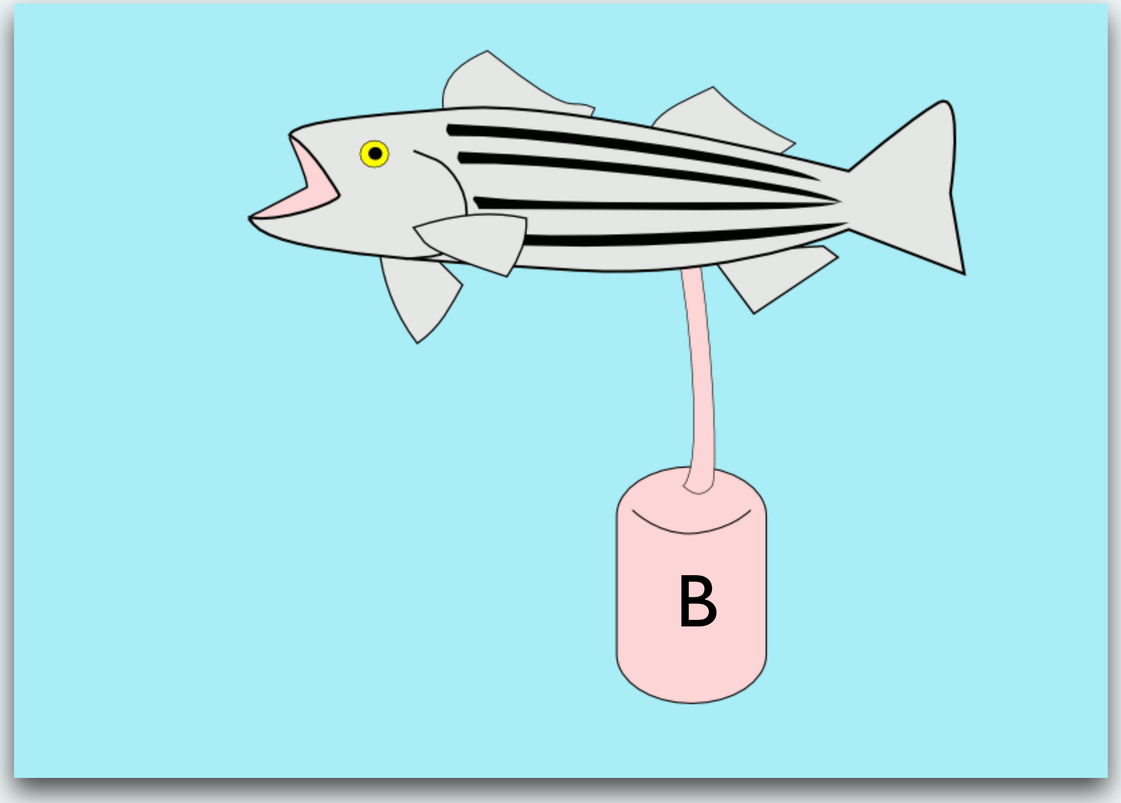




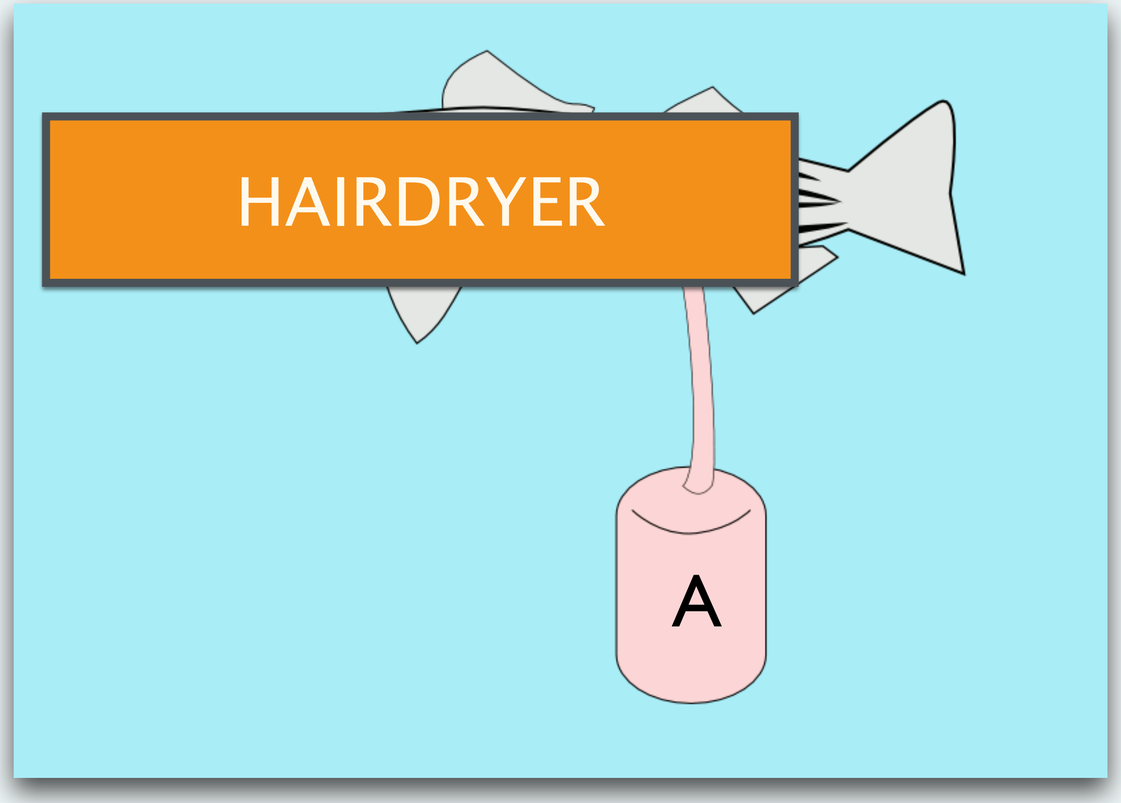
HAIRDRYER

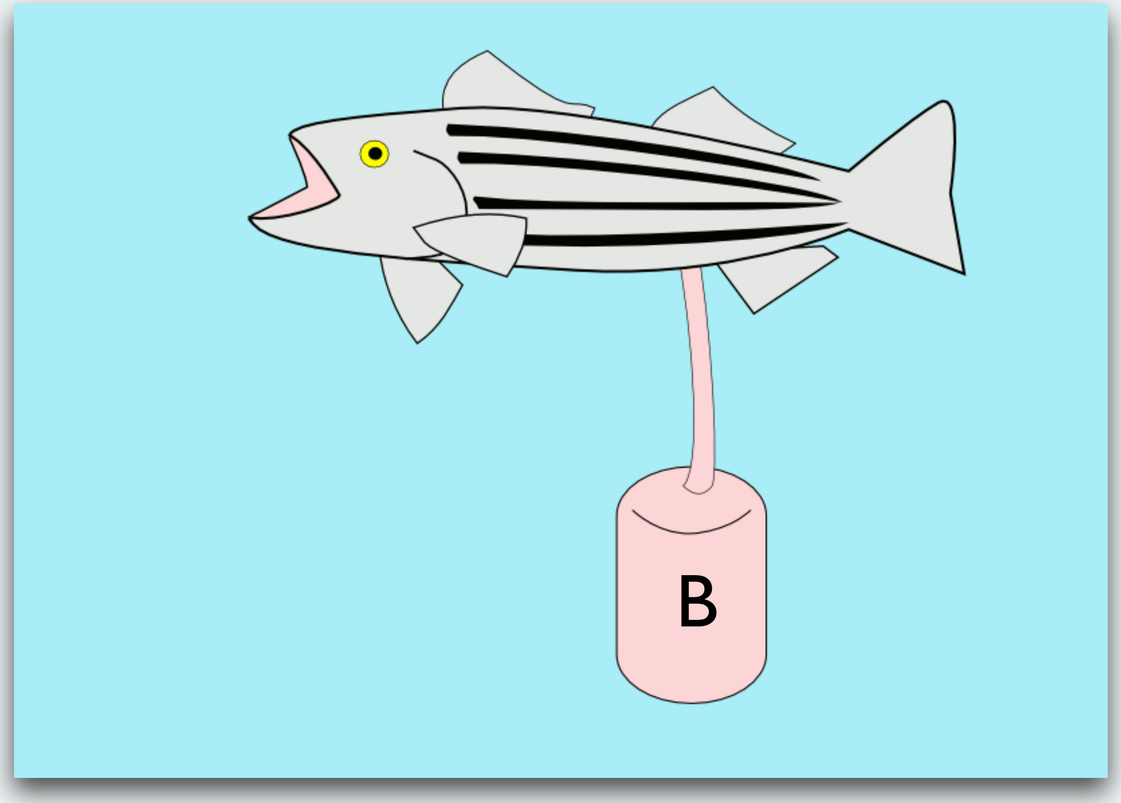
riak





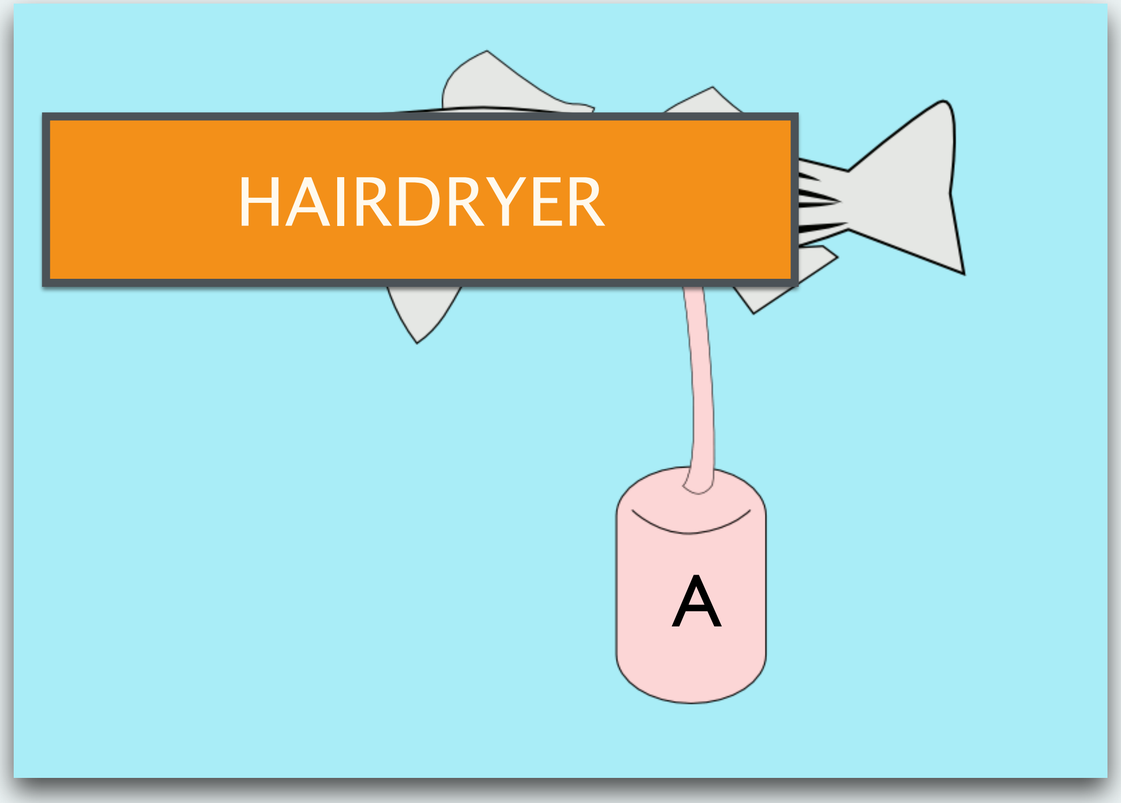
riak

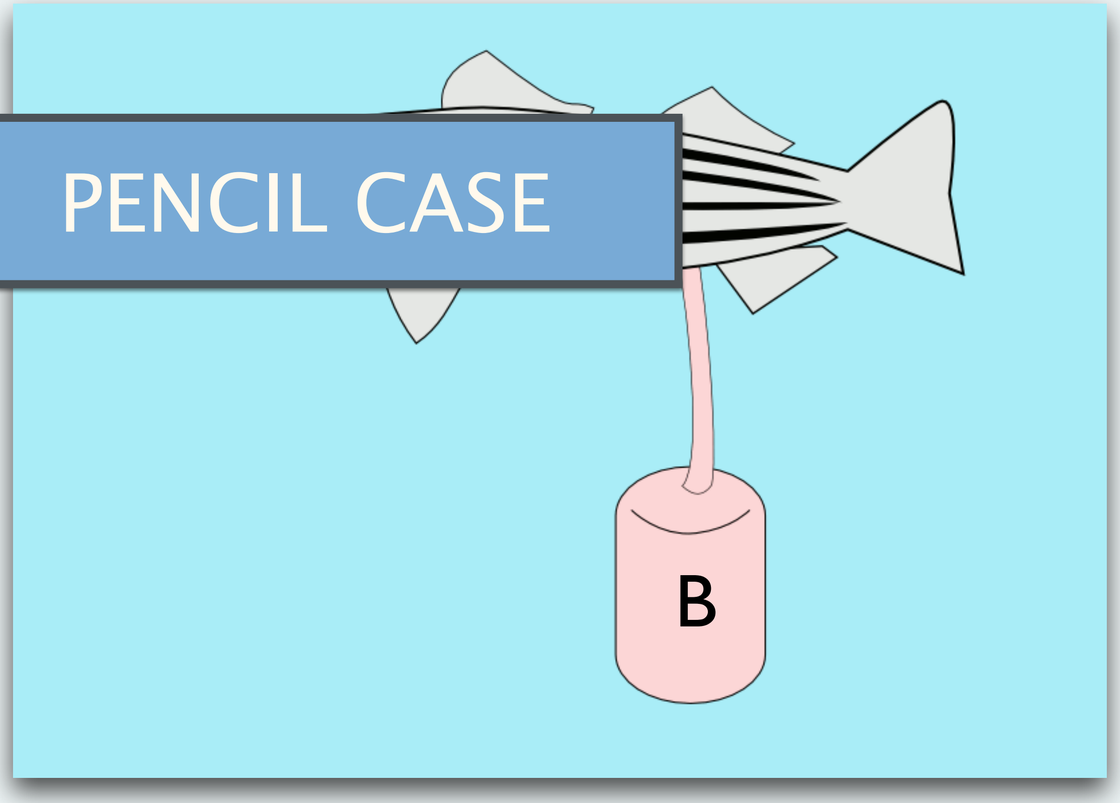




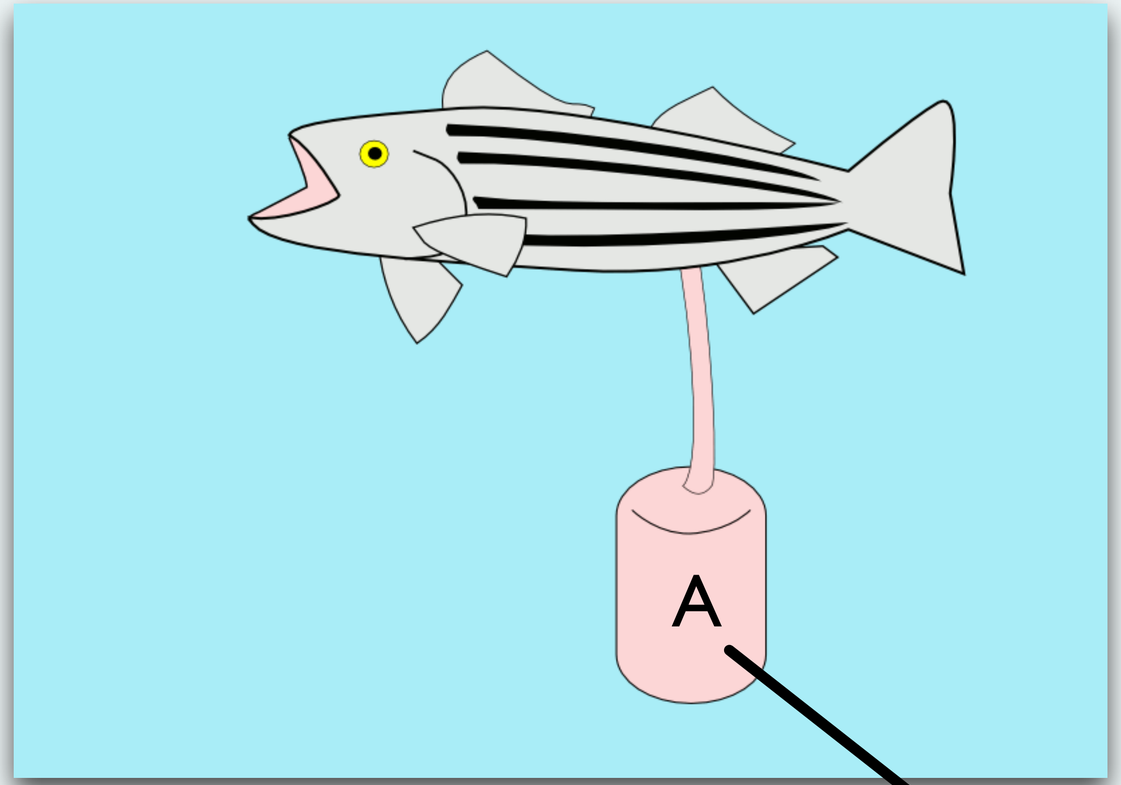
PENCIL CASE

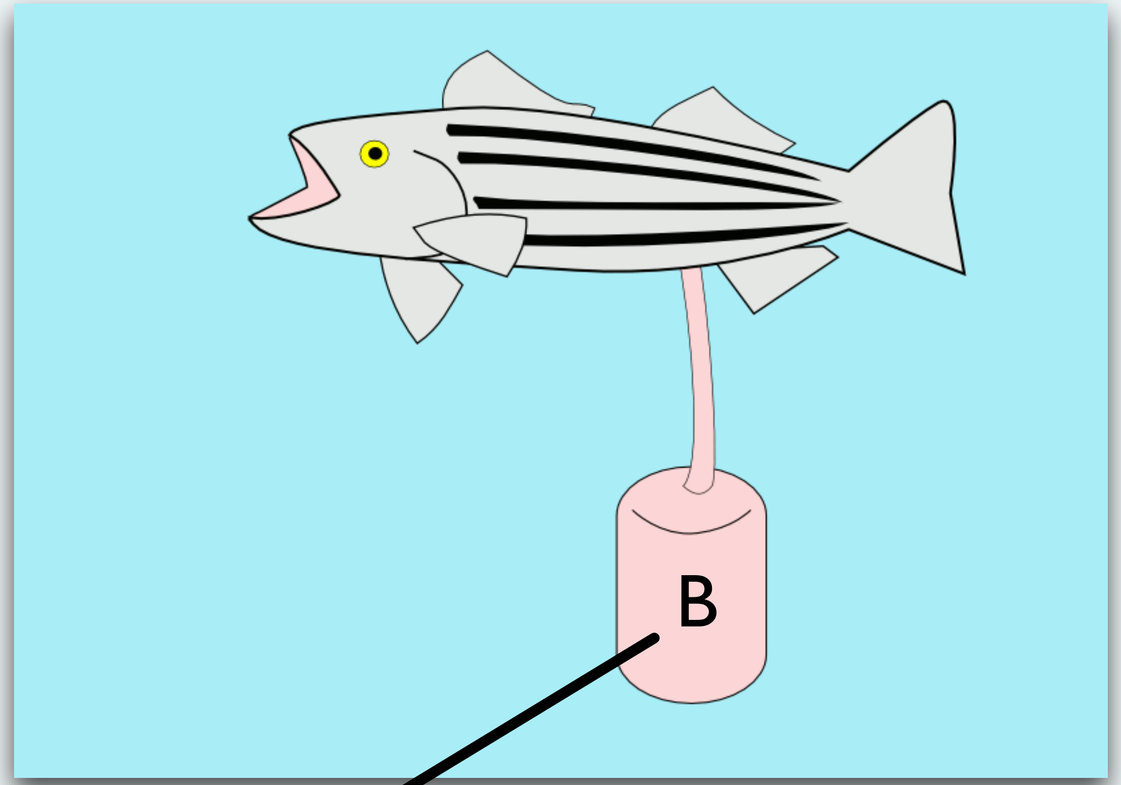
riak





riak





[HAIRDRYER], [PENCIL CASE]

riak

Merge

Set Union of Values
Simples, right?

Merge

Deterministic

Merge

Deterministic
Idempotent

Merge

Deterministic

Idempotent

Associative

Merge

Deterministic
Idempotent
Associative
Commutative

Removes?

Set Union?

Absence

How can you tell if X is missing from A and not B because A hasn't seen the addition, or if A has removed X?

Complexity

Ad Hoc

CRDTs

CRDTs

Convergent Replicated Data Types

CRDTs

Commutative Replicated Data Types

CRDTs

Conflict-free Replicated Data Types

Theory

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A comprehensive study of Convergent and Commutative Replicated Data Types

Marc Shapiro, INRIA & LIP6, Paris, France
Nuno Preguiça, CITI, Universidade Nova de Lisboa, Portugal
Carlos Baquero, Universidade do Minho, Portugal

Marek Zawirski, INRIA & UPMC, Paris, France

13 Jan 2011

riak

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

An Optimized Conflict-free Replicated Set

Annette Bieniusa, INRIA & UPMC, Paris, France
Marek Zawirski, INRIA & UPMC, Paris, France
Nuno Preguiça, CITI, Universidade Nova de Lisboa, Portugal
Marc Shapiro, INRIA & LIP6, Paris, France
Carlos Baquero, HASLab, INESC TEC & Universidade do Minho, Portugal
Valter Balegas, CITI, Universidade Nova de Lisboa, Portugal

Sérgio Duarte CITI, Universidade Nova de Lisboa, Portugal

11 Oct 2012

riak

Dotted Version Vectors: Logical Clocks for Optimistic Replication

Nuno Preguiça
CITI/DI
FCT, Universidade Nova de Lisboa
Monte da Caparica, Portugal
nmp@di.fct.unl.pt

Carlos Baquero, Paulo Sérgio Almeida,
Victor Fonte, Ricardo Gonçalves
CCTC/DI
Universidade do Minho
Braga, Portugal
{cbm,psa,vff}@di.uminho.pt, rtg@lsd.di.uminho.pt

Abstract

In cloud computing environments, a large number of users access data stored in highly available storage systems. To provide good performance to geographically disperse users and allow operation even in the presence of failures or network partitions, these systems often rely on optimistic replication solutions that guarantee only eventual consistency. In this scenario, it is important to be able to accurately and efficiently

The mentioned systems follow a design where the data store is always writable. A consequence is that replicas of the same data item are allowed to diverge, and this divergence should later be repaired. Accurate tracking of concurrent data updates can be achieved by a careful use of well established causality tracking mechanisms [5], [6], [7], [8]. In particular, for data storage systems, version vectors [6] enables the system to compare any pair of replica versions and detect if

Bounded Join Semilattices

Bounded Join Semilattices

Partially ordered set; least upper bound; ACI.

Bounded Join Semilattices

Associativity: (X Y) Z = X (Y Z)

Bounded Join Semilattices

Commutativity: X Y = Y X

Bounded Join Semilattices

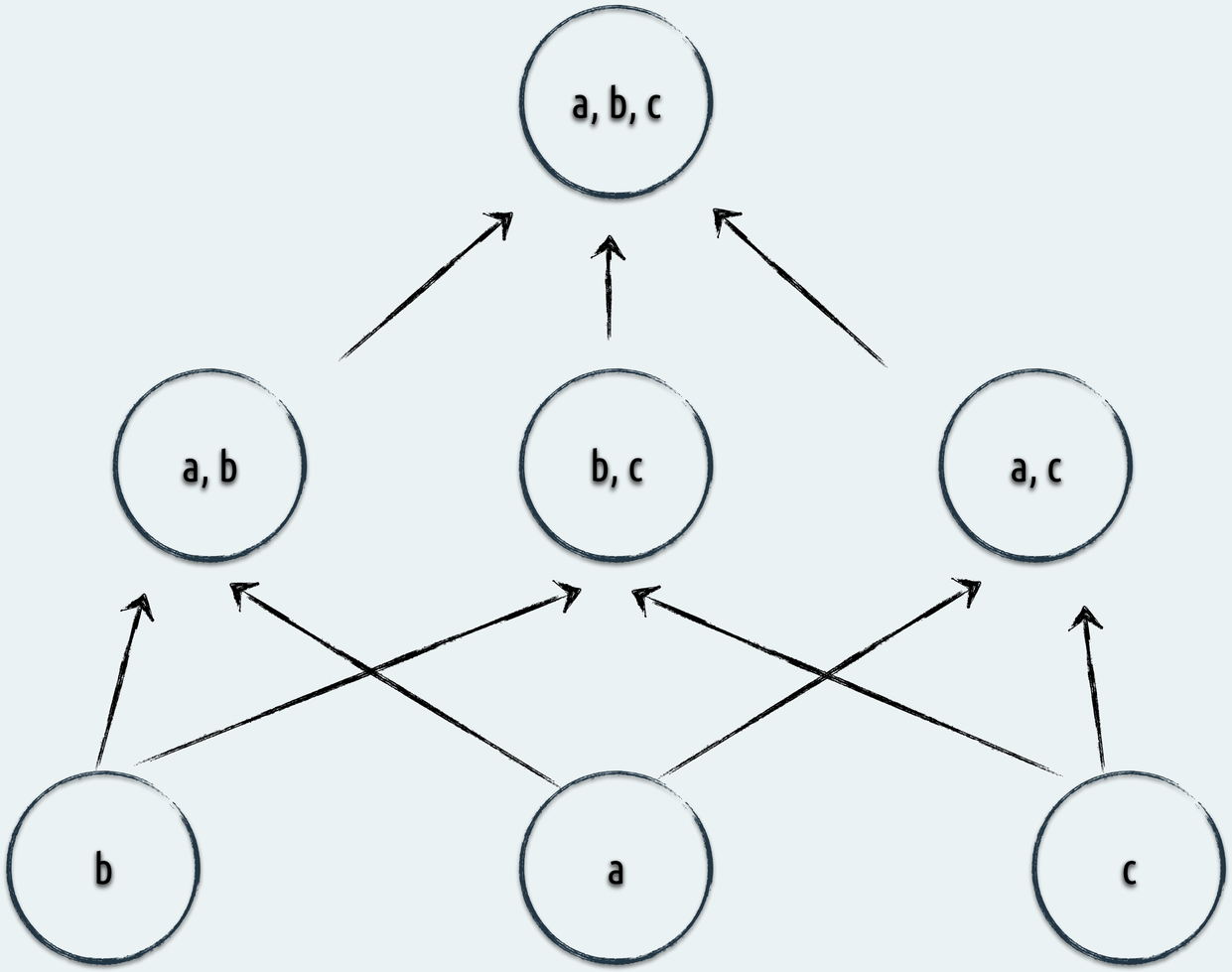
Idempotence: X X = X

Bounded Join Semilattices

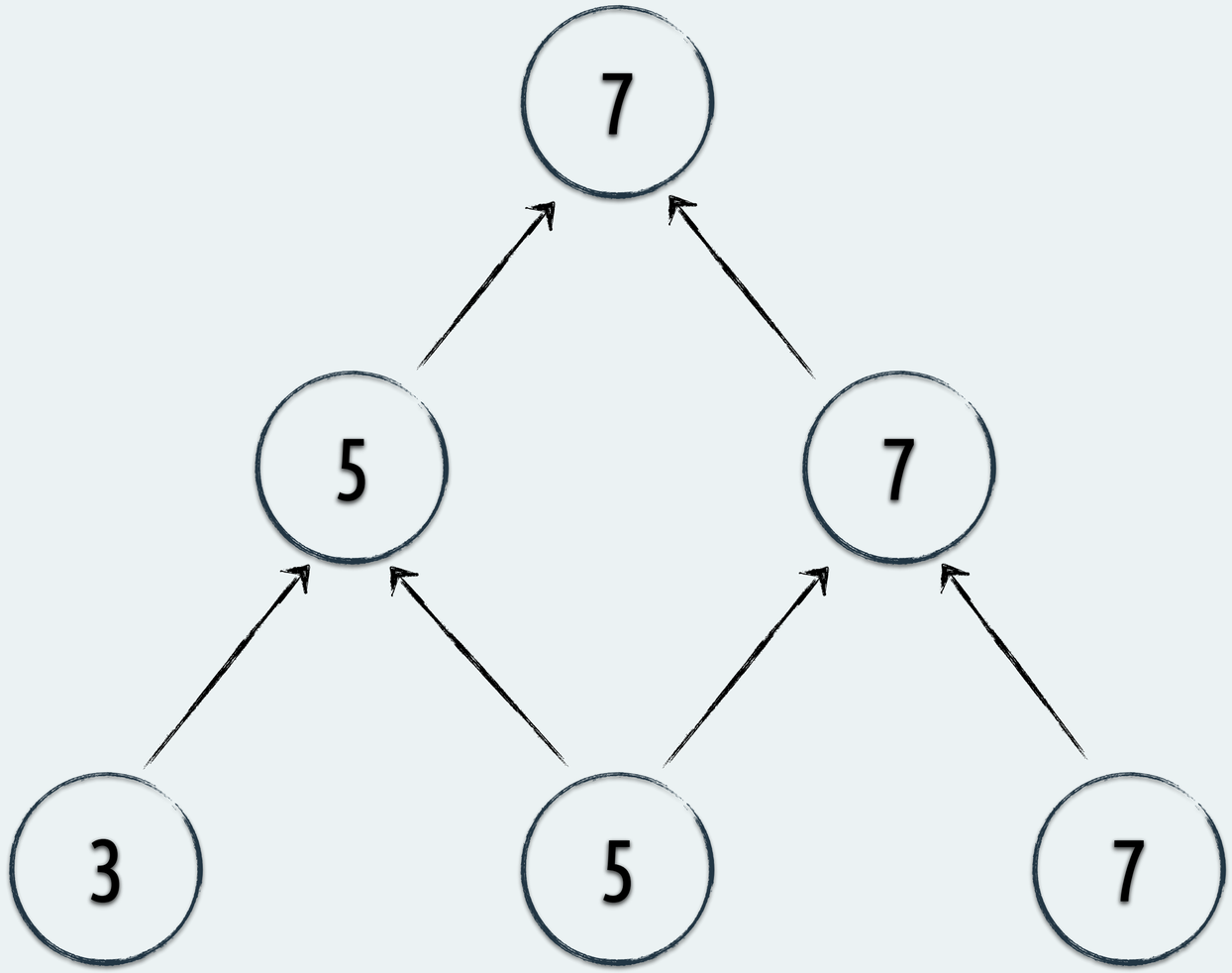
Objects grow over time; merge computes LUB

Bounded Join Semilattices

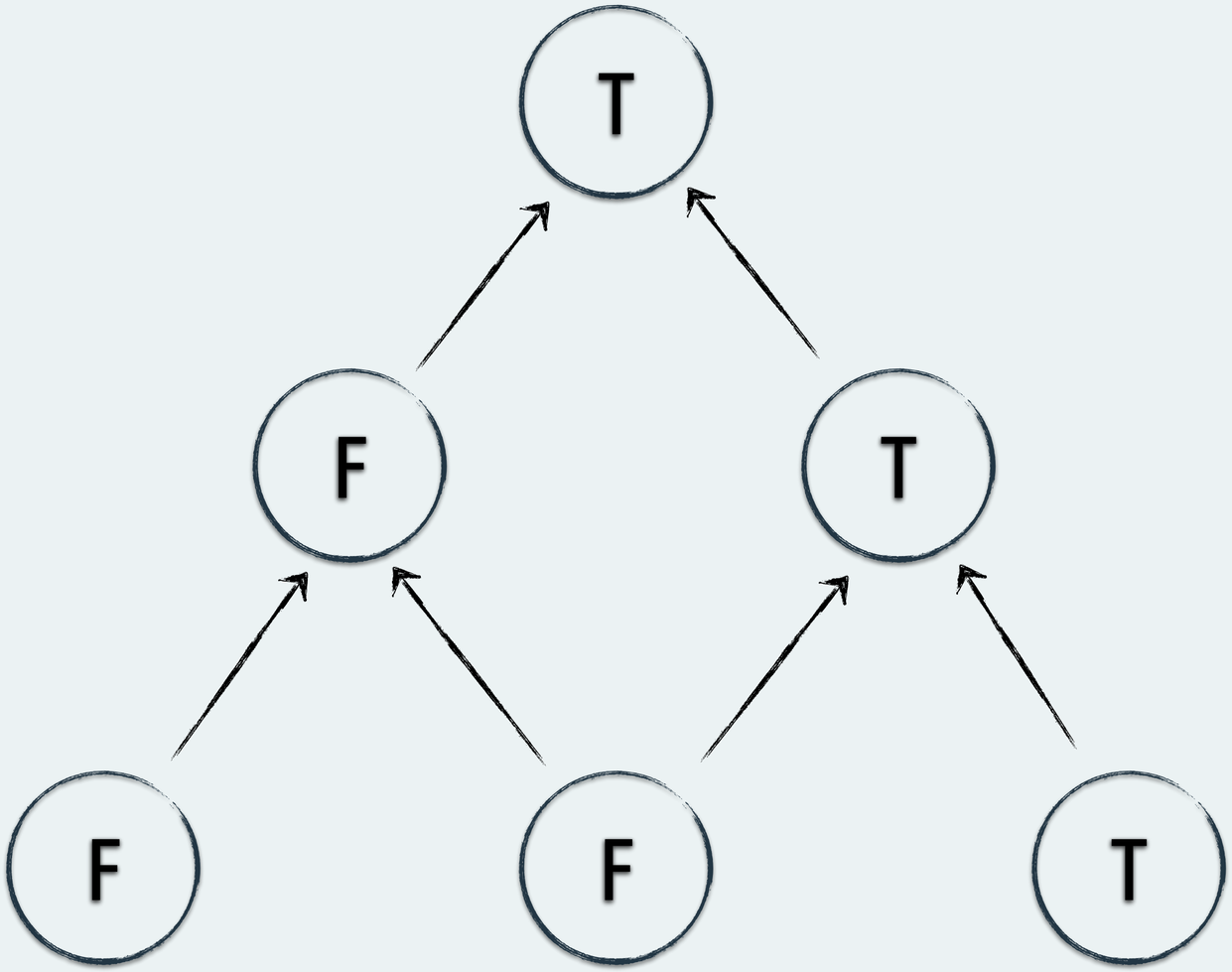
Examples



Set; merge function: union.



Increasing natural; merge function: max.



Booleans; merge function: or.

Merge

Deterministic
Idempotent
Associative
Commutative

Riak 2.0

Pick your semantic

Add wins
Remove wins
Keep both

Trade Off

More metadata == bigger objects

Actors?

Version Vectors
Entry Per Actor (Charron-Bost)

Riak 1.4

Counters: non-idempotent; 0 (Actors)

Riak 2.0

Sets: Add, Remove, Membership; Idempotent

Riak 2.0

Sets: Add wins 0(Actors + Elements)

Riak 2.0

Maps: Recursive; Associative Array; Nestable

Riak 2.0

Maps: Update wins; 0(Actors + Elements)

Riak 2.0

Maps: LWW-Register, Booleans, Sets and Maps

Riak 2.0

LWW-Register: last writer wins

Riak 2.0

Boolean: Enabled, Disabled; 0(Actors)

riak_dt

git clone git@github.com:basho/riak_dt.git

Evolution

of a Set

Causality

Version Vectors

[{a, 1}, {b, 3}, {c, 2}]

Causality

Version Vectors

[{a, 2}, {b, 3}, {c, 2}] > [{a, 1}, {b, 3}, {c, 2}]

Causality

Version Vectors

[{a, 2}, {b, 3}, {c, 2}]

[{a, 1}, {b, 4}, {c, 2}]

[{a, 2}, {d, 1}, {c, 2}]

[{a, 2}, {b, 4}, {c, 2}]

Causality
Version Vectors

‘Dots’ are Events

Causality

‘Dots’ are Events

Evolution of a Set

G-SET

Evolution of a Set

G-SET

Evolution of a Set

G-SET

2P-SET

Evolution of a Set

U-SET

Evolution of a Set

U-SET

OR-SET

Evolution of a Set

U-SET

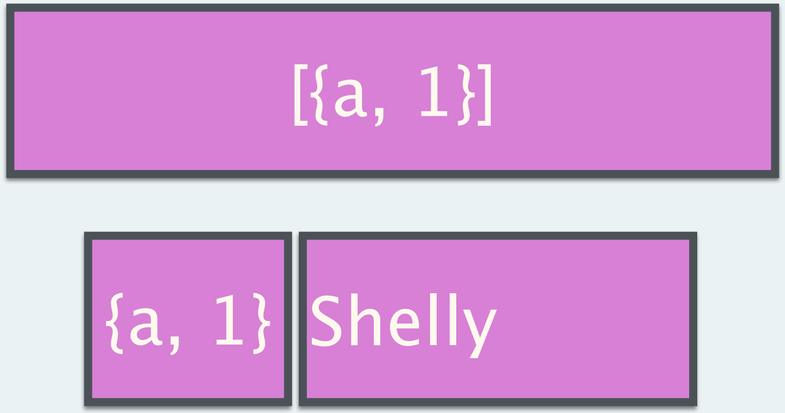
OR-SET

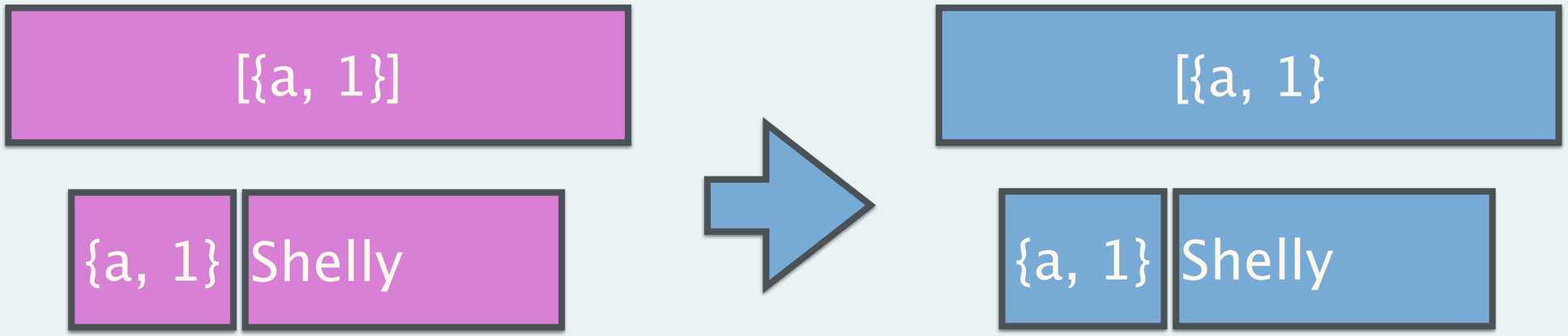
Evolution of a Set

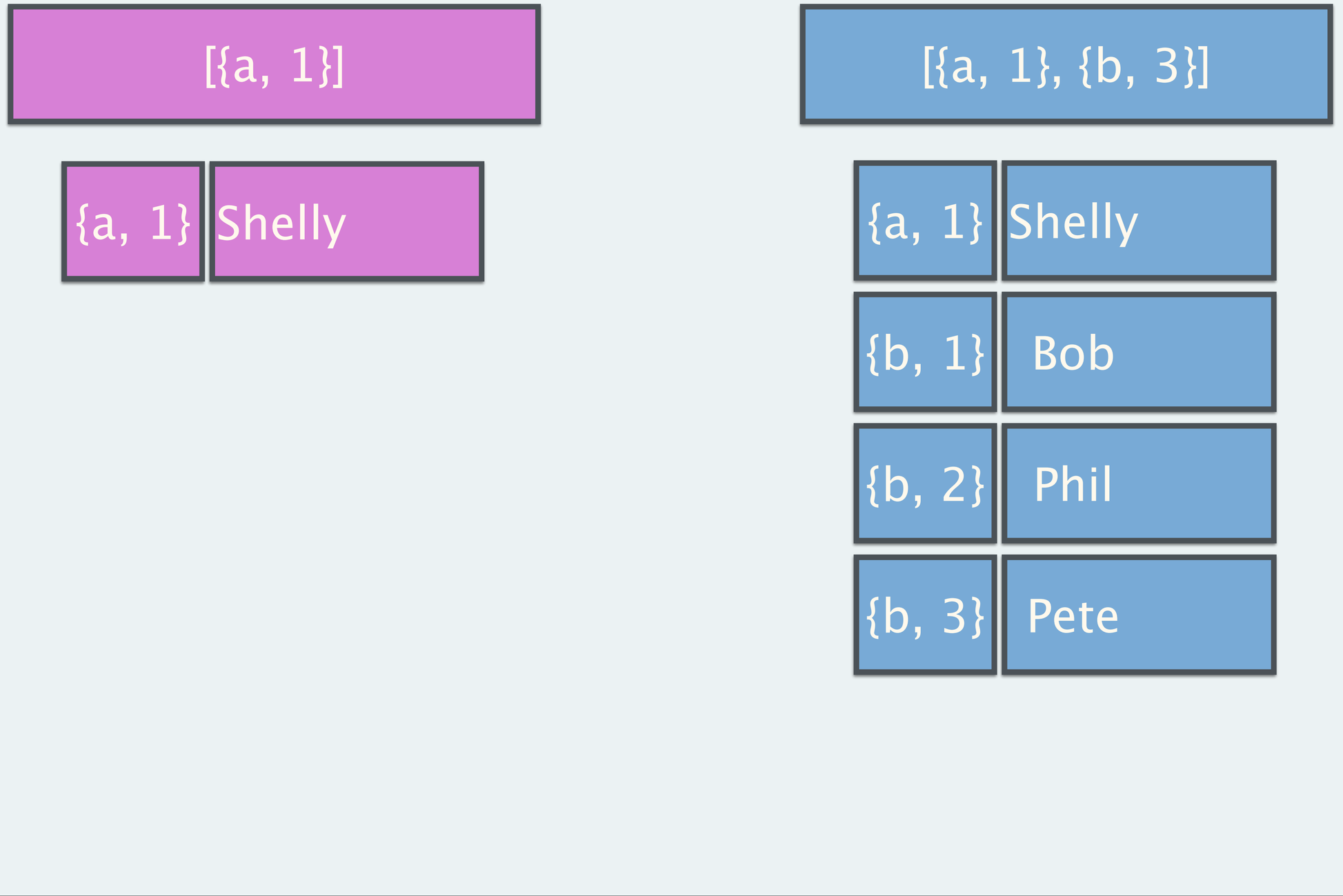
U-SET

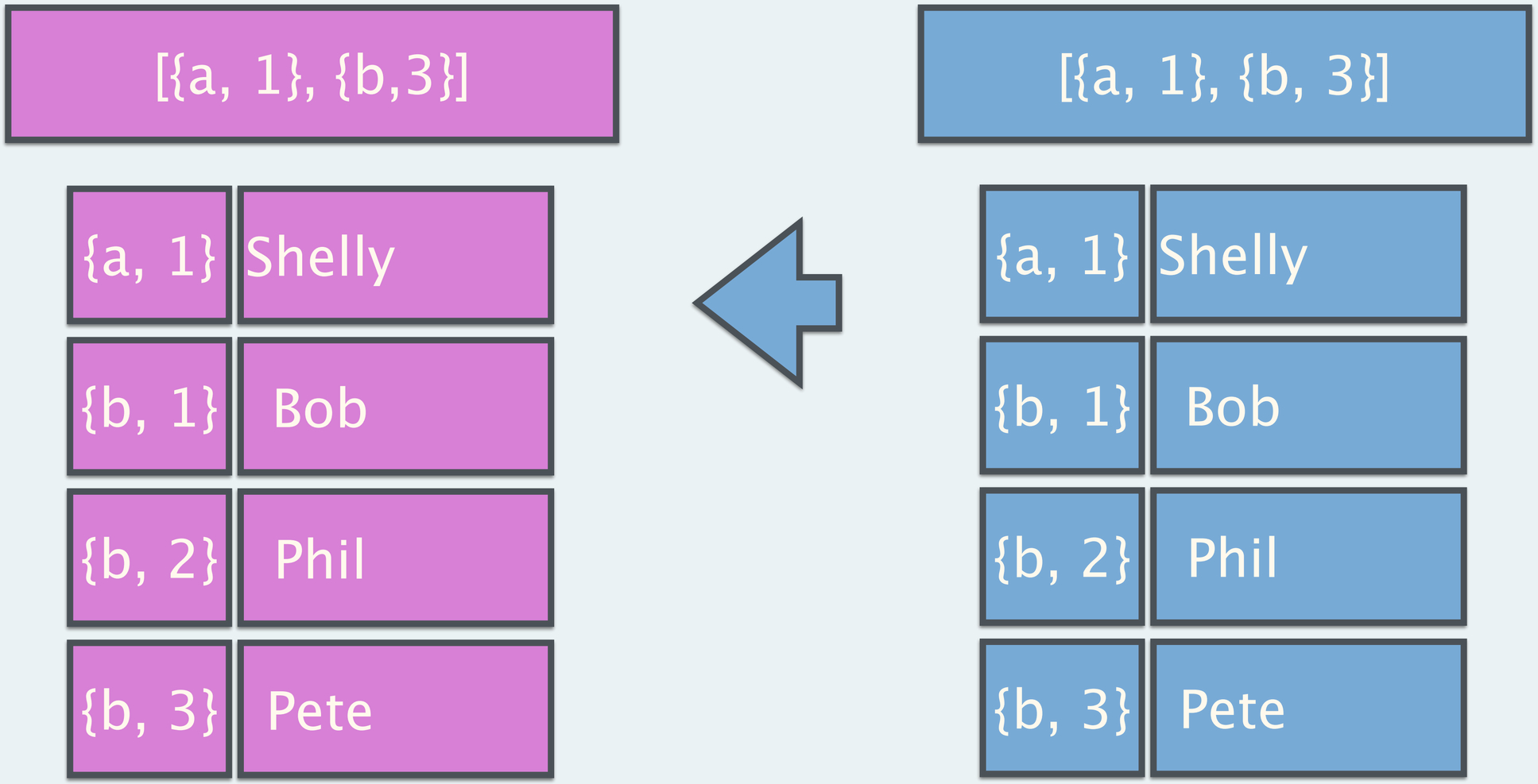
OR-SET

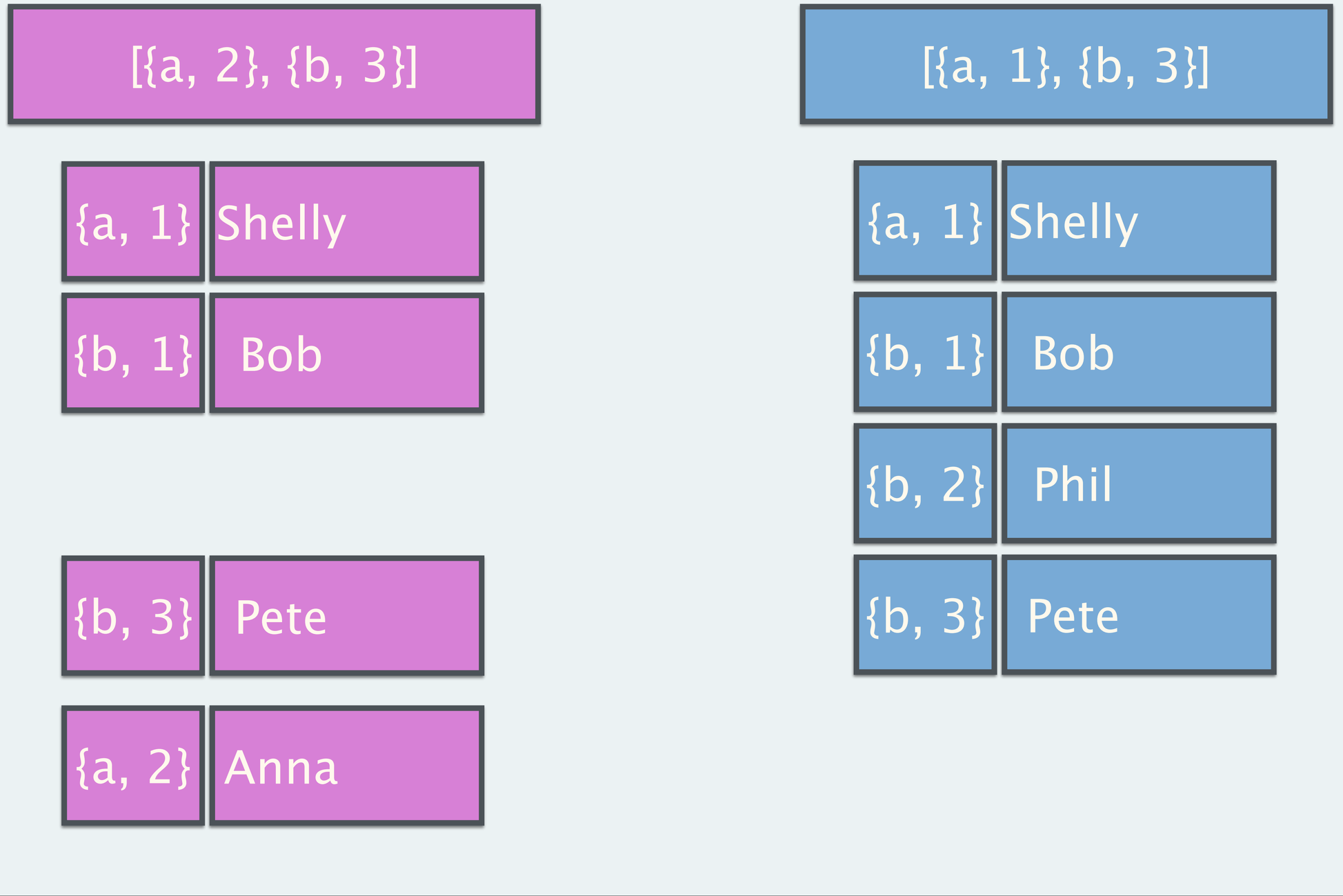
OR-SWOT

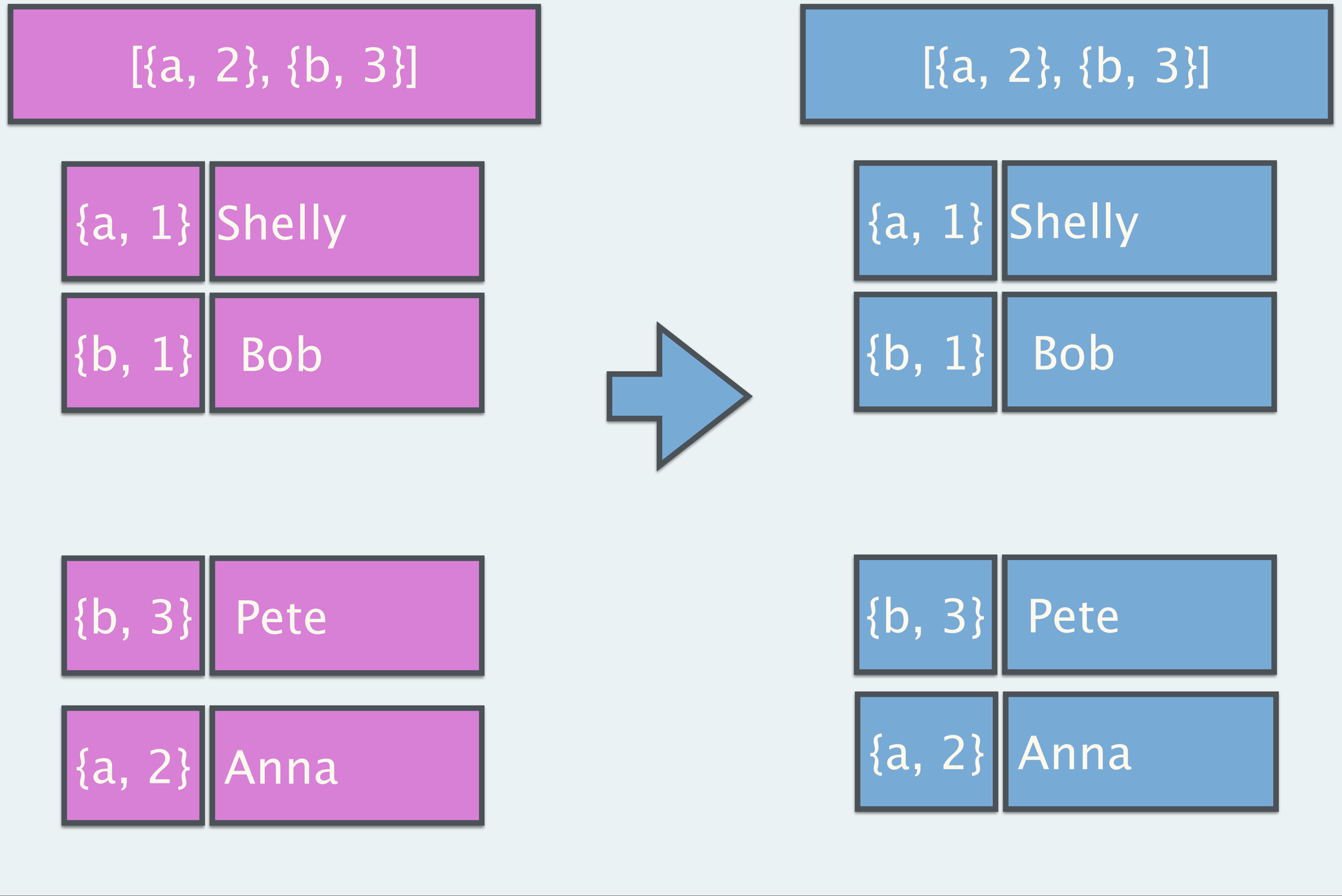

[{a, 1}]
{a, 1}
Shelly


[{a, 1}]
{a, 1}
Shelly
[{a, 1}
{a, 1}
Shelly


[{a, 1}]
{a, 1}
Shelly
[{a, 1}, {b, 3}]
{a, 1}
Shelly
{b, 1}
Bob
{b, 2}
Phil
{b, 3}
Pete


[{a, 1}, {b,3}]
{a, 1}
Shelly
{b, 1}
Bob
{b, 2}
Phil
{b, 3}
Pete
[{a, 1}, {b, 3}]
{a, 1}
Shelly
{b, 1}
Bob
{b, 2}
Phil
{b, 3}
Pete


[{a, 2}, {b, 3}]
{a, 1}
Shelly
{b, 1}
Bob
{b, 3}
Pete
{a, 2}
Anna
[{a, 1}, {b, 3}]
{a, 1}
Shelly
{b, 1}
Bob
{b, 2}
Phil
{b, 3}
Pete


[{a, 2}, {b, 3}]
{a, 1}
Shelly
{b, 1}
Bob
{b, 3}
Pete
{a, 2}
Anna
[{a, 2}, {b, 3}]
{a, 1}
Shelly
{b, 1}
Bob
{b, 3}
Pete
{a, 2}
Anna

GARBAGE

Tombstones
Dead Actors

GARBAGE

~~Tombstones~~

Dead Actors

No More Siblings!


SyncFree – Home
https://syncfree.lip6.fr
SyncFree
ABOUT
NEWS
CONSORTIUM
EVENTS
PUBLICATIONS
Large-scale computation without synchronisation
Large-scale on-line services including social networks and multiplayer games handle huge quantities of frequently changing shared data. Maintaining its consistency is relatively simple in a centralised cloud, but no longer possible due to increased scalability requirements. Instead, data must replicated across several distributed data centres, requiring new principled approaches to consistency that will be explored by the SyncFree project.
SyncFree is a European research project taking place for 3 years, staring October 2013, and is funded by the European Union, grant agreement n°609551.
Featured
SyncFree at Cloudscape
Members of the SyncFree project attended the Cloudscape VI even held in Brussels on February 24th-25th. This series of workshops brings together public and private sector IT and legal professionals to discuss current and future issues and opportunities presented by the cloud.
A position paper about SyncFree published at the workshop can be found at the following link:
http://admin.cloudscapeseries.eu/Repository/document/PositionP

Riak 2.0 Dev Preview

http://docs.basho.com/riak/2.0.0pre11/downloads/

Bucket Types

Strong Consistency

Security

Search 2.0

basho

WELCOME TO

BASHO CAREERS

We make some of the most **powerful** stuff in the world. And we want your help.

http://bashojobs.resumator.com

riak

Questions?

@JoelJacobson

joel@basho.com