Apache Cassandra

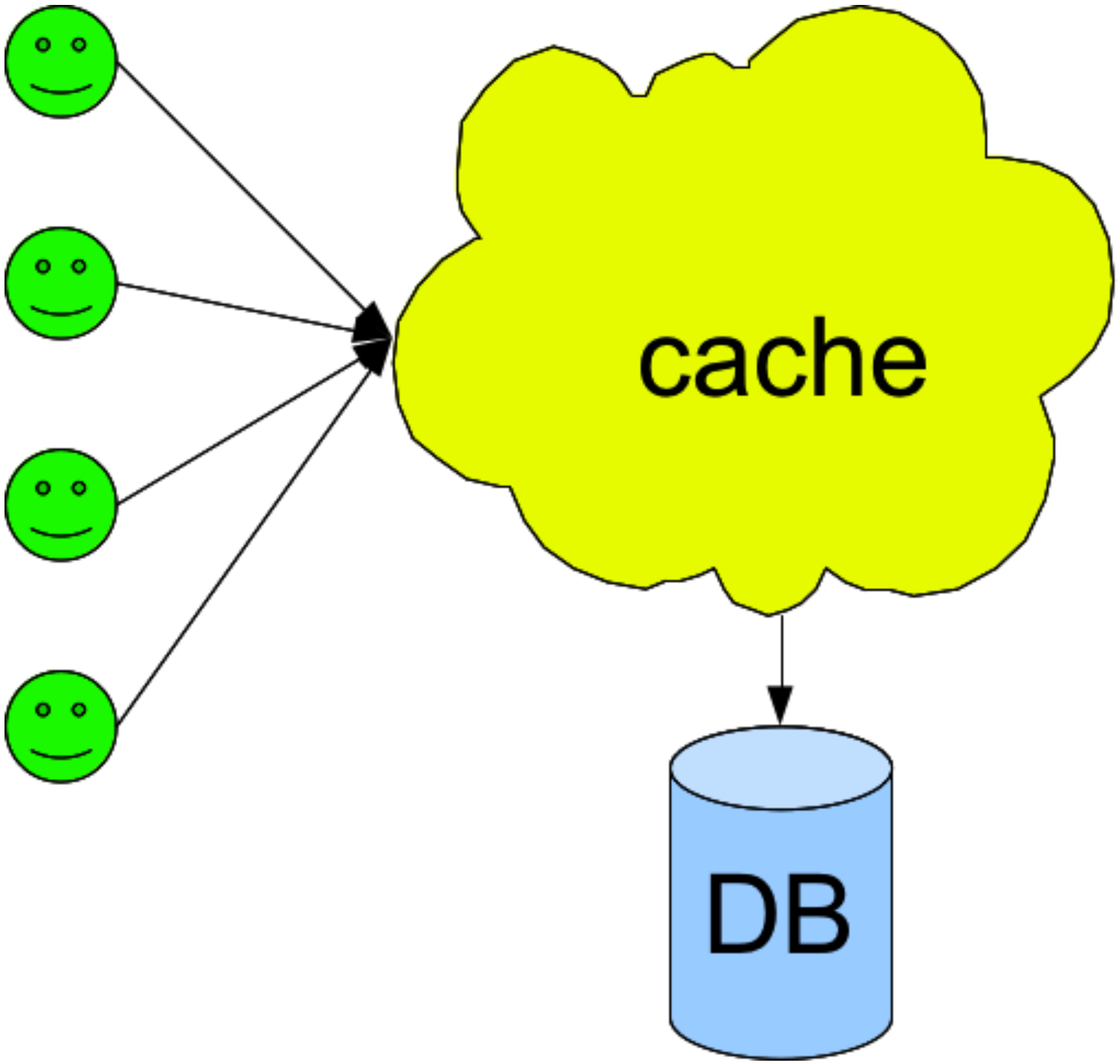
Jonathan Ellis

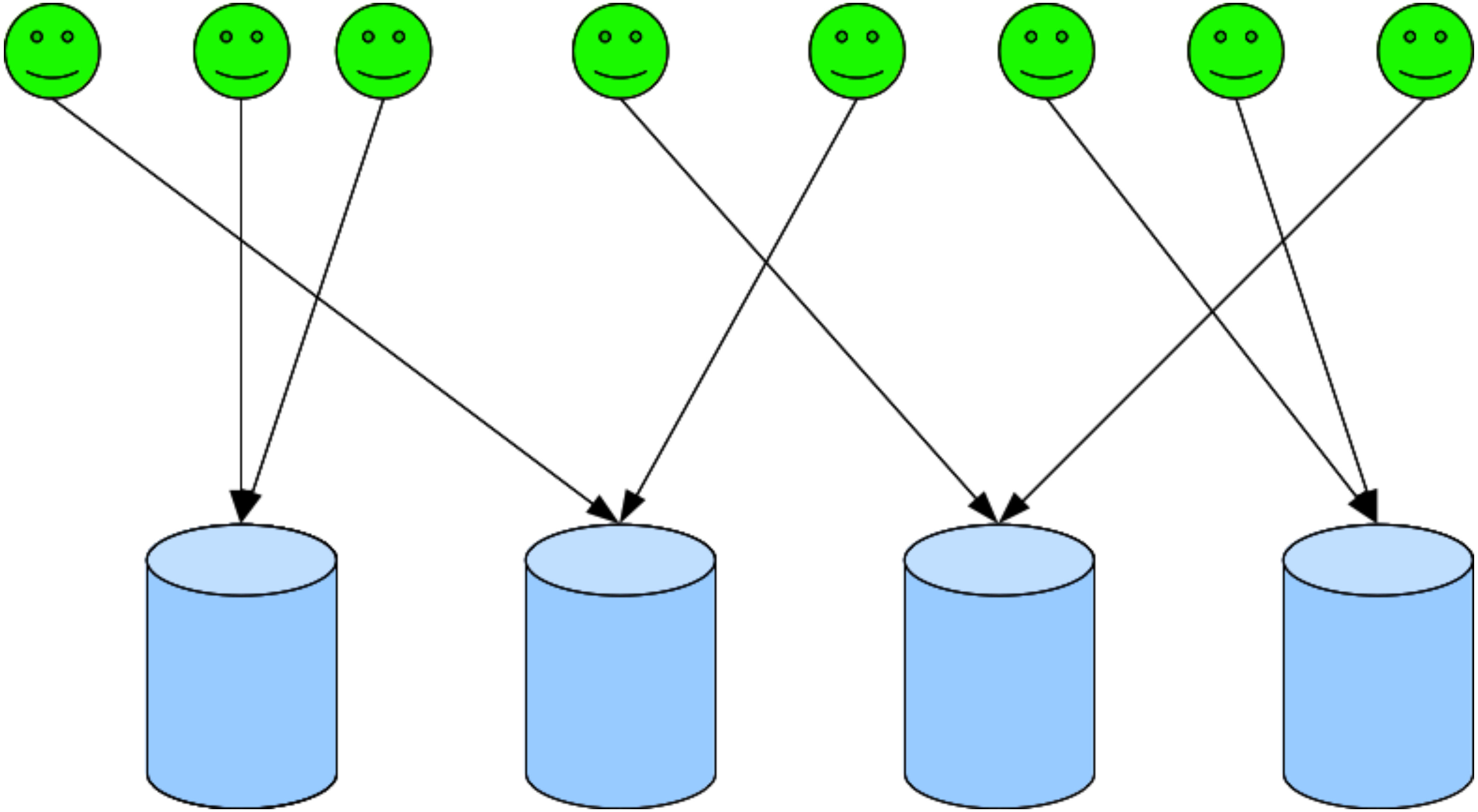
jbellis@riptano.com / @spyced

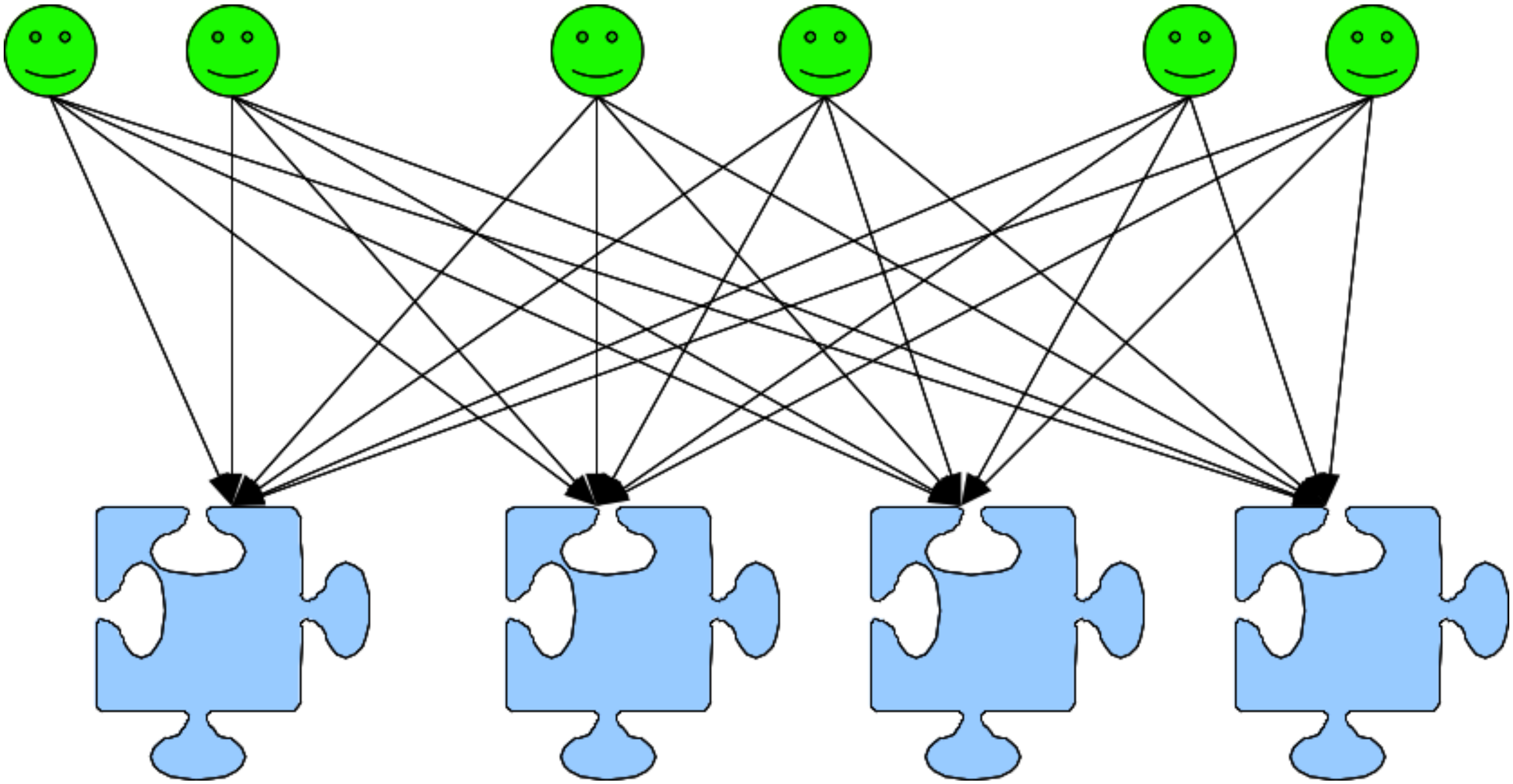
Why Cassandra?

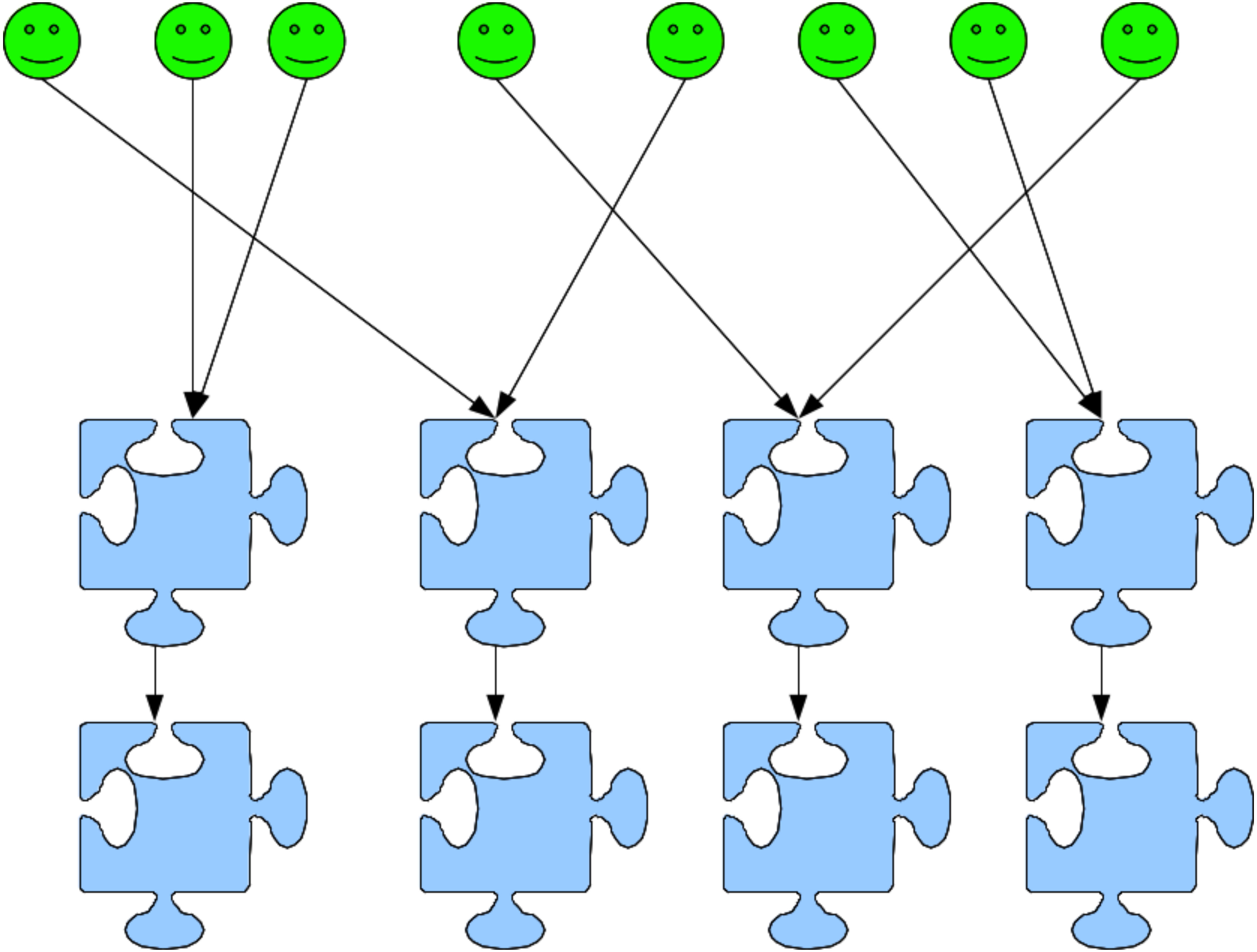
- Relational databases are not designed to scale
- B-trees are slow
  - and require read-before-write

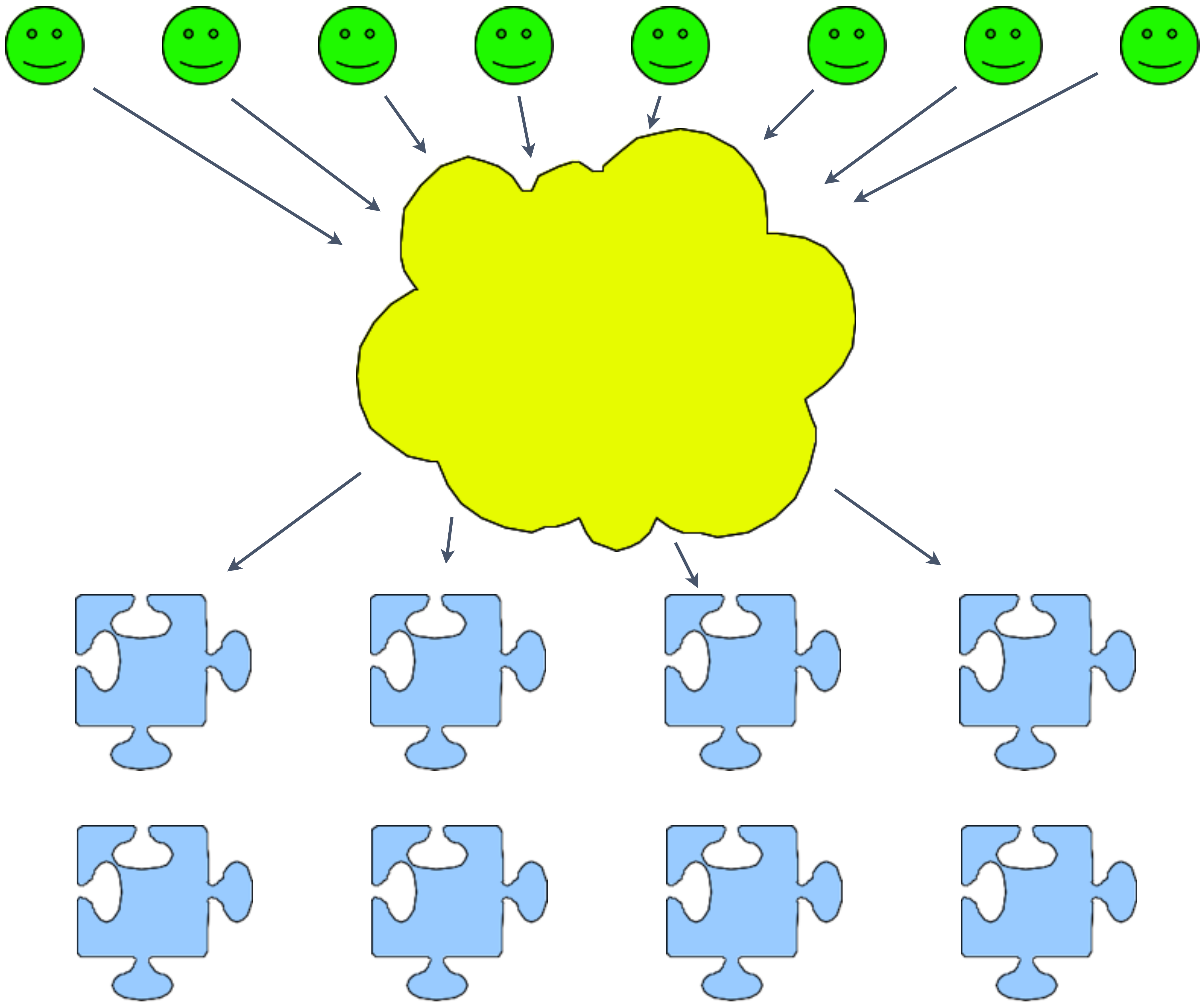
Sun microsystems
Sun microsystems

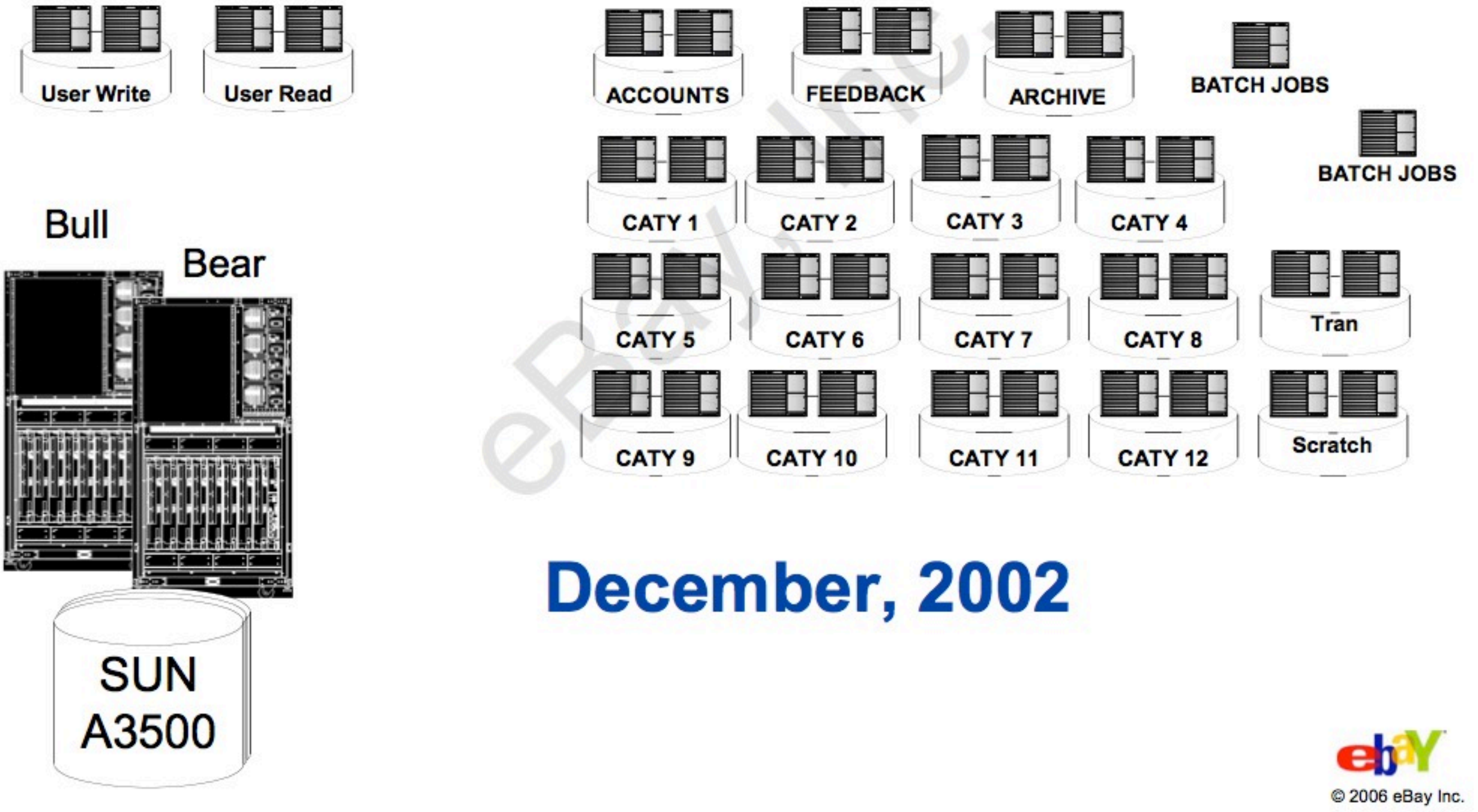

cache
DB









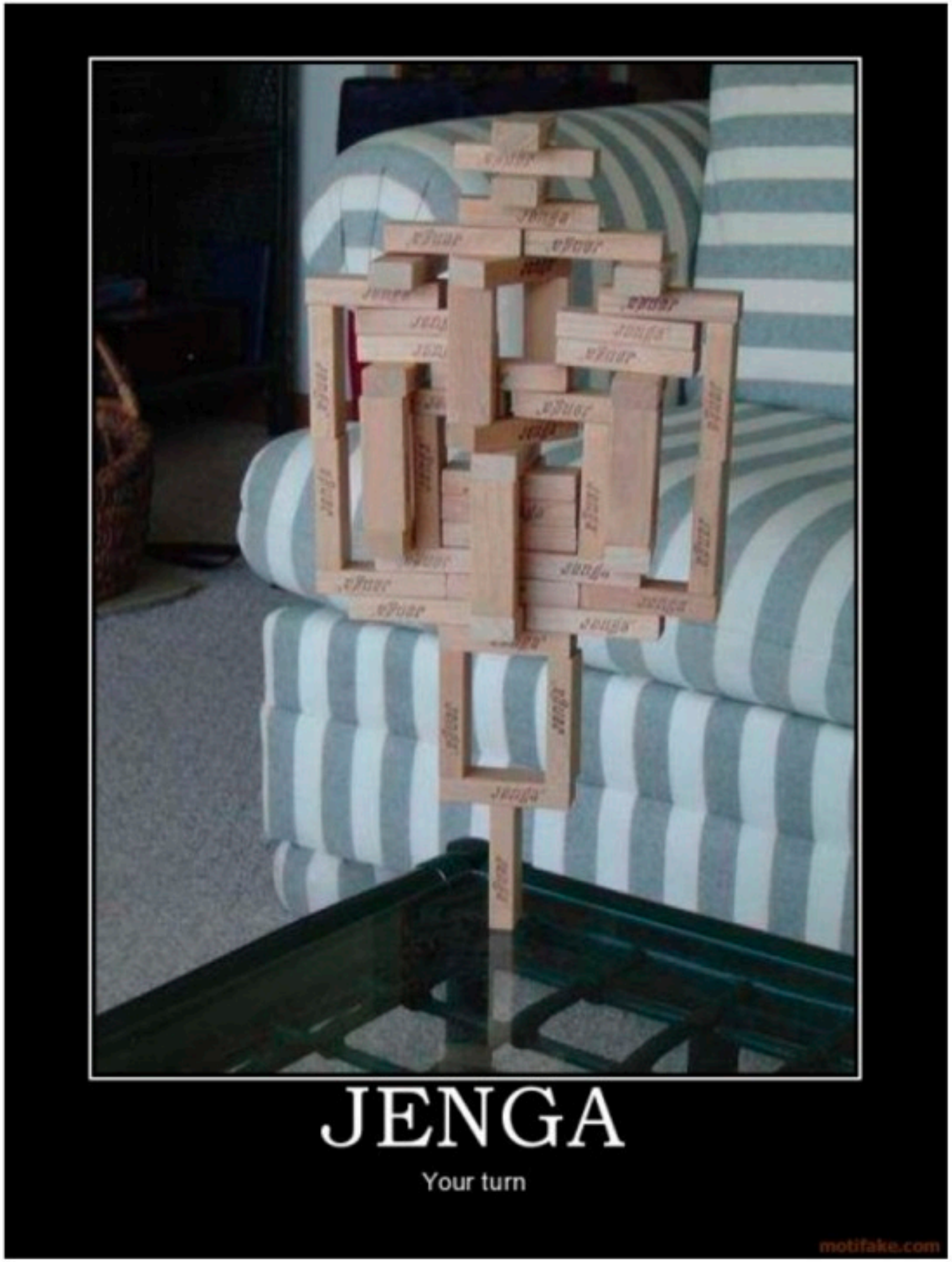


eBay, Inc.

December, 2002

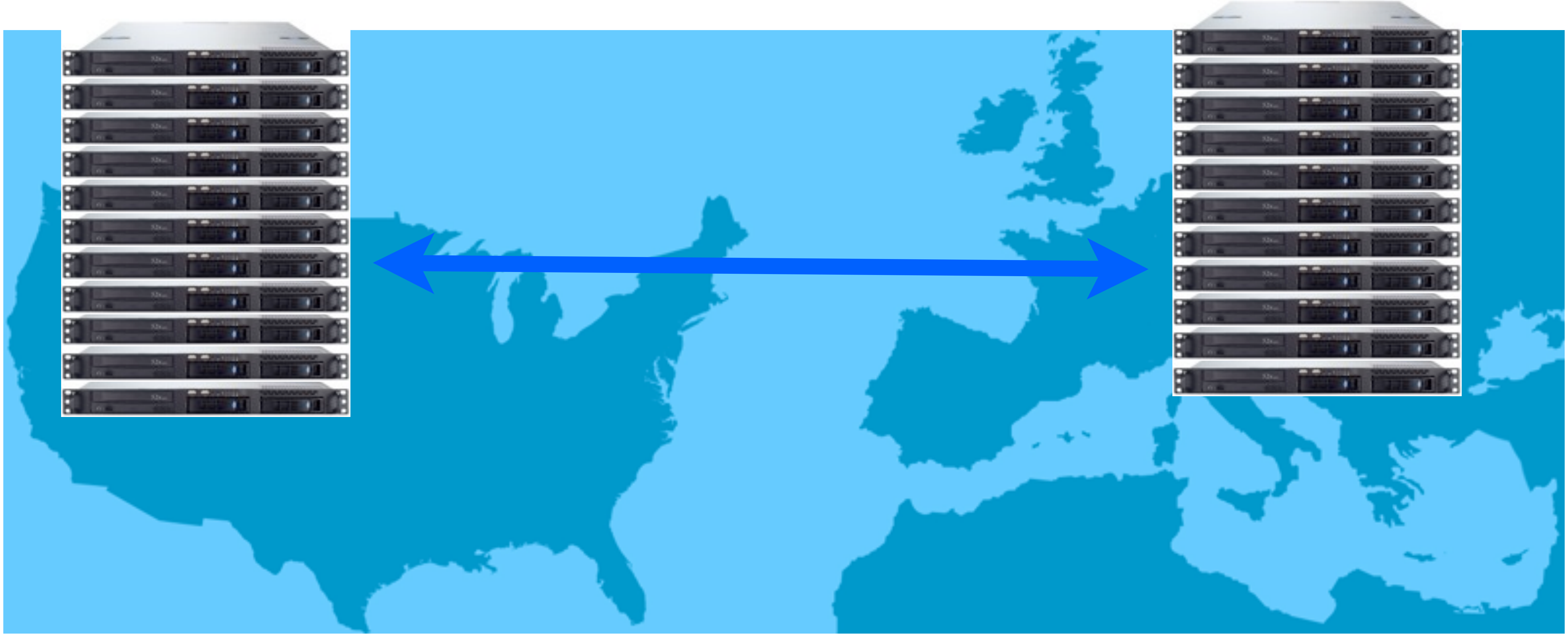
© 2006 eBay Inc.

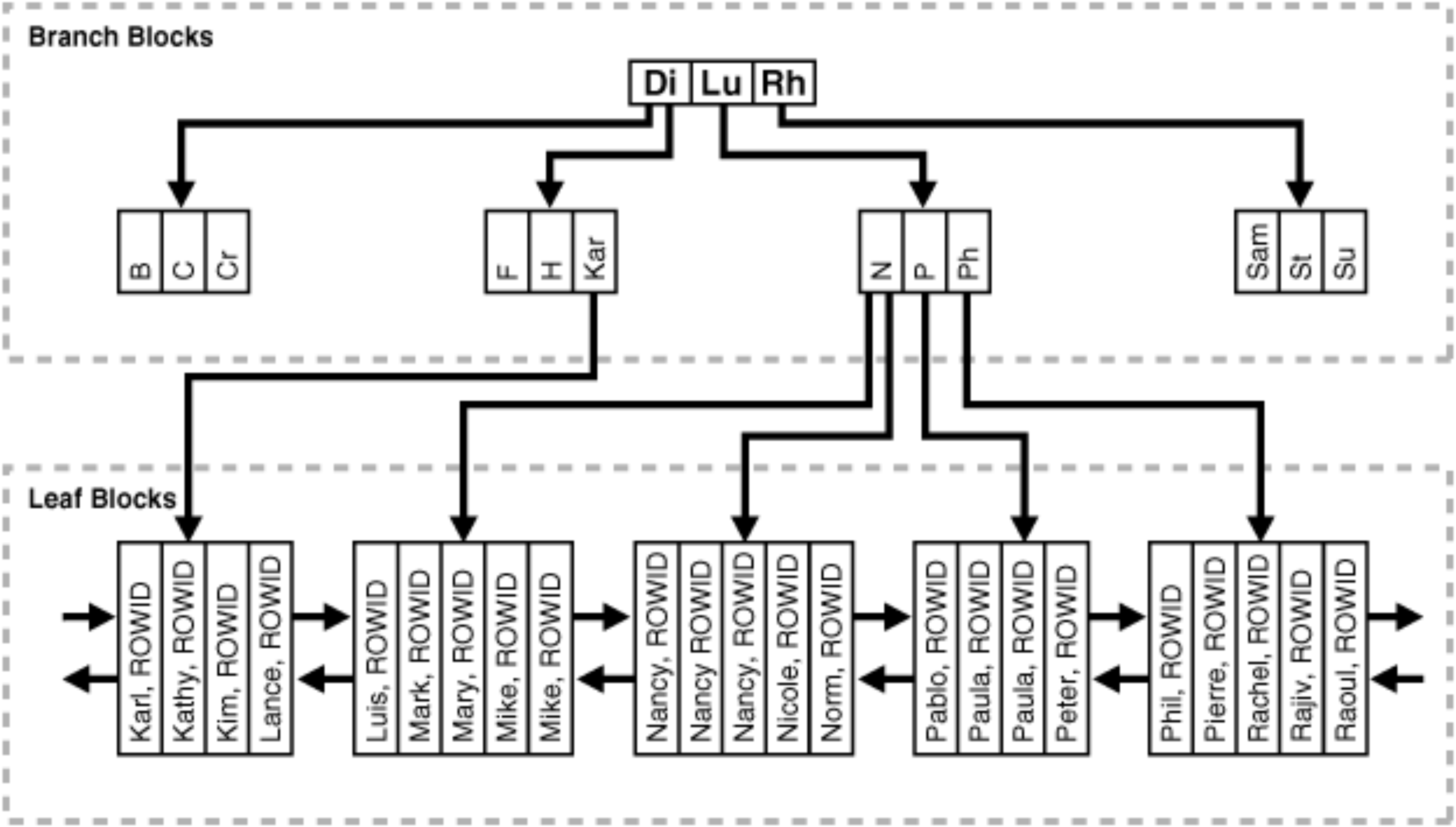
(“The eBay Architecture,” Randy Shoup and Dan Pritchett)


JENGA
Your turn
motifake.com

eBay: NoSQL pioneer

- “BASE is diametrically opposed to ACID. Where ACID is pessimistic and forces consistency at the end of every operation, BASE is optimistic and accepts that the database consistency will be in a state of flux. Although this sounds impossible to cope with, in reality it is quite manageable and leads to levels of scalability that cannot be obtained with ACID.”
  - **”BASE: An Acid Alternative,” Dan Pritchett, eBay**





Branch Blocks

Di Lu Rh

B C Cr

F H Kar

N P Ph

Sam St Su

Leaf Blocks

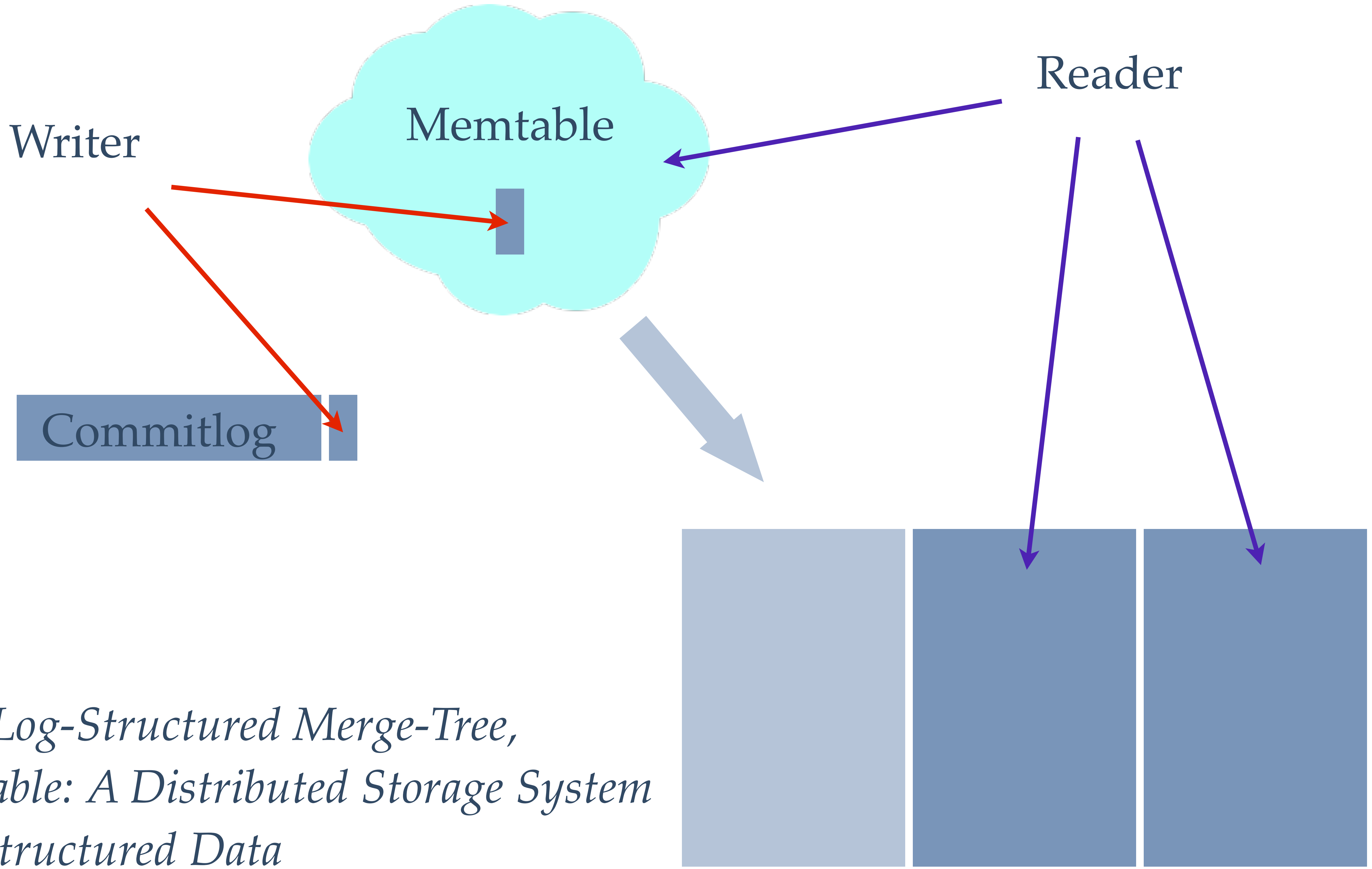
Karl, ROWID
Kathy, ROWID
Kim, ROWID
Lance, ROWID

Luis, ROWID
Mark, ROWID
Mary, ROWID
Mike, ROWID
Mike, ROWID

Nancy, ROWID
Nancy ROWID
Nancy, ROWID
Nicole, ROWID
Norm, ROWID

Pablo, ROWID
Paula, ROWID
Paula, ROWID
Peter, ROWID

Phil, ROWID
Pierre, ROWID
Rachel, ROWID
Rajiv, ROWID
Raoul, ROWID



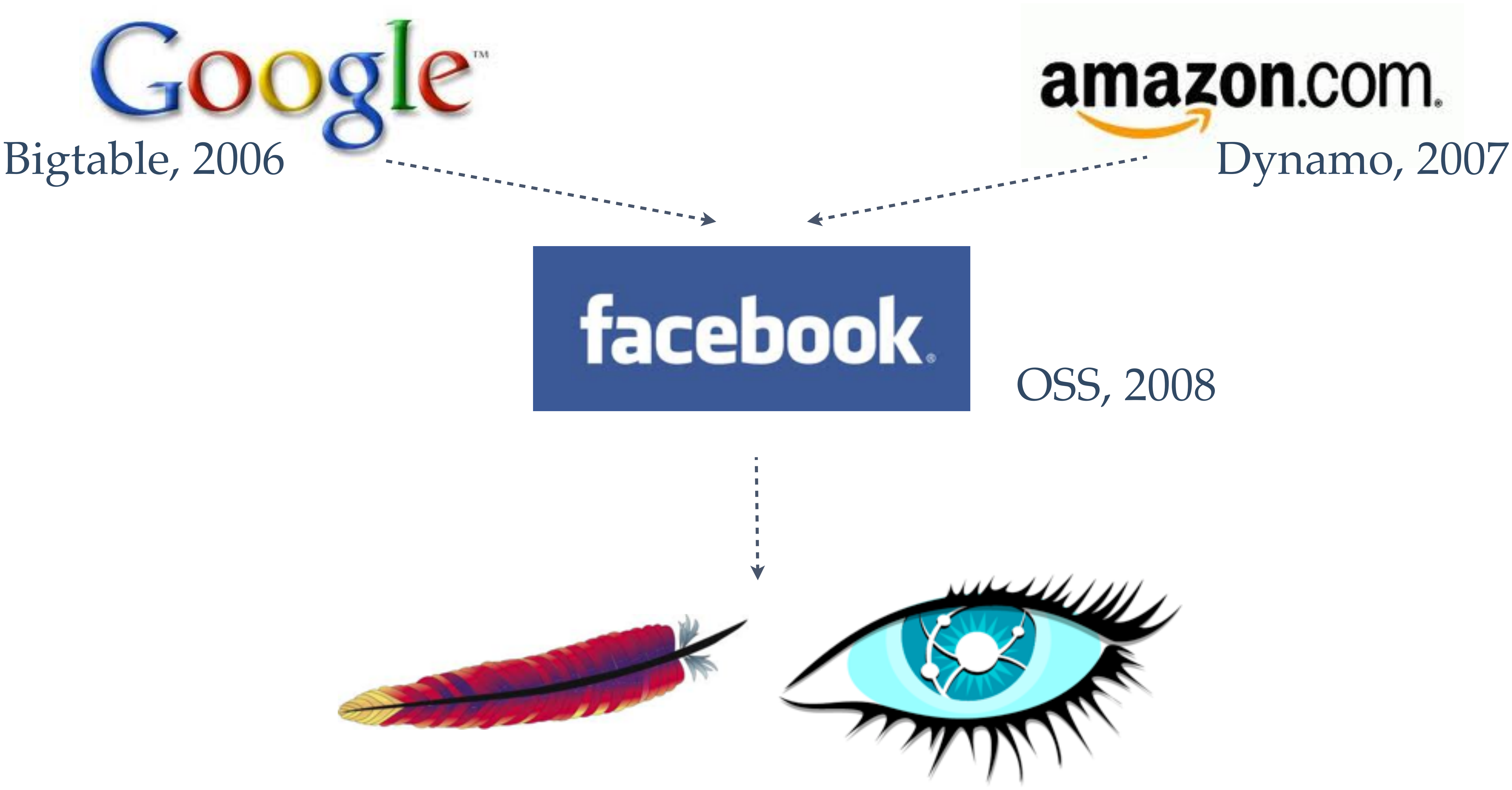
Writer

Memtable

Reader

Commitlog

The Log-Structured Merge-Tree,
Bigtable: A Distributed Storage System for Structured Data



Bigtable, 2006

Dynamo, 2007

OSS, 2008

Incubator, 2009

TLP, 2010

Myth 1

- “NoSQL is for people who don’t understand {SQL, denormalization, query tuning, ...}”
  - Similarly: “Only users of [database X] are turning to NoSQL databases, because X sucks.”

Myth 2

- “NoSQL is nothing new because we had key / value databases like bdb years ago.”

The downside to NoSQL-as-identifier

Myth 3

- “Only huge sites like Facebook and Twitter need to care about scalability.”

Cassandra in production

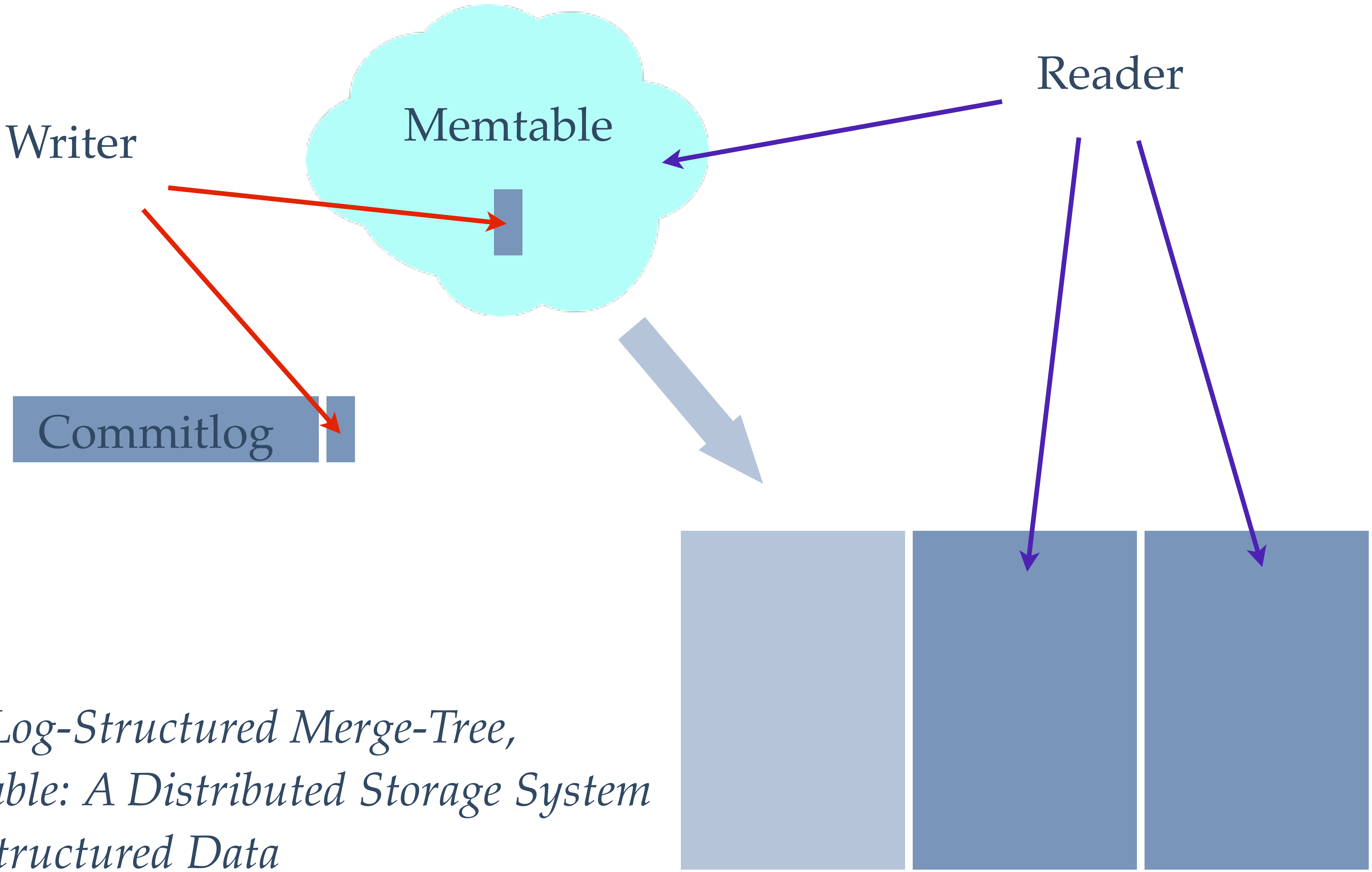
- Digital Reasoning: NLP + entity analytics
- OpenX: largest publisher-side ad network in the world
- Cloudkick: performance data & aggregation
- SimpleGEO: location-as-API
- Ooyala: video analytics and business intelligence
- ngmoco: massively multiplayer game worlds

Myth 4

- NoSQL is only appropriate for unimportant data

Durabilty

- Write to commitlog
  - fsync is cheap since it's append-only
- Write to memtable
- [amortized] flush memtable to sstable



Writer

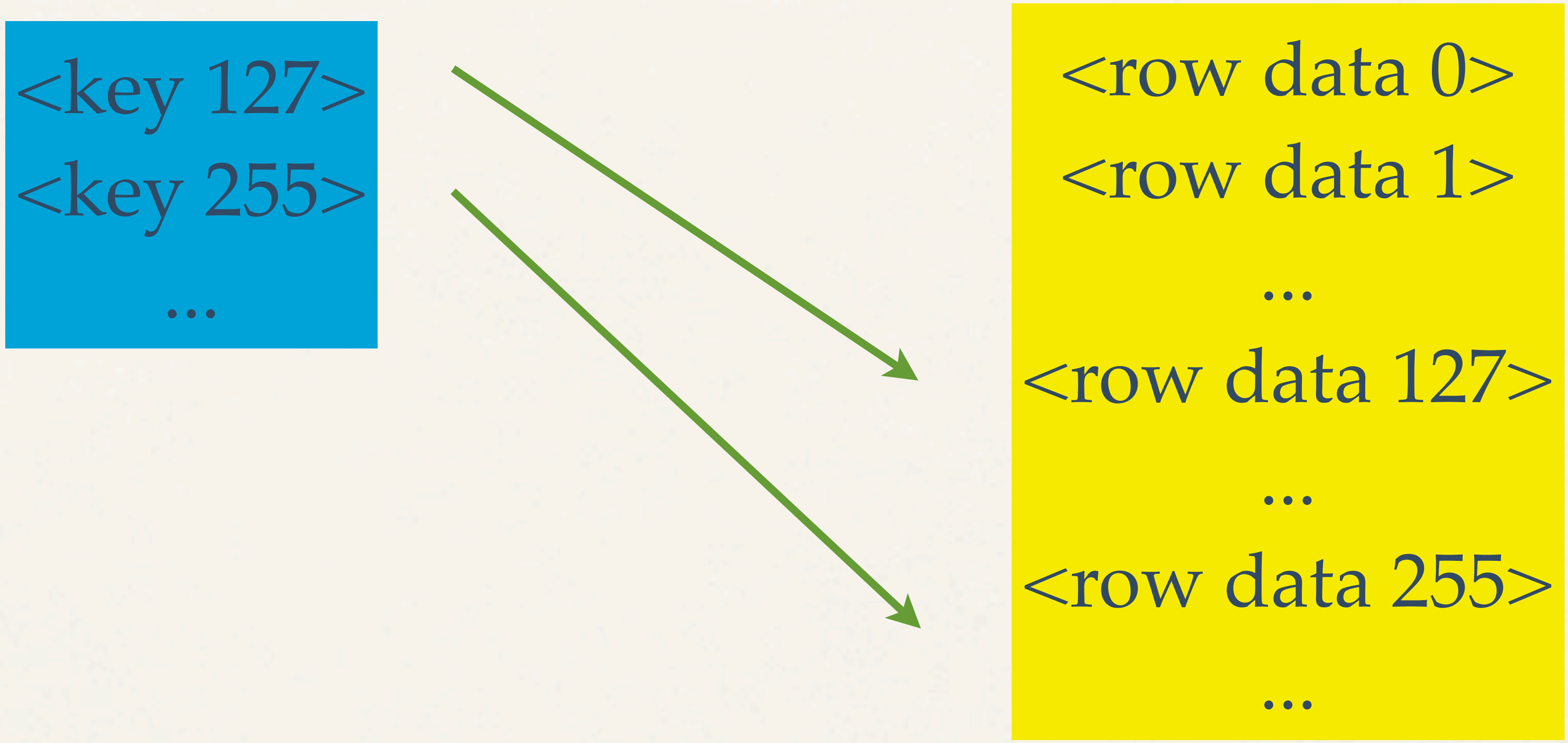
Memtable

Reader

Commitlog

The Log-Structured Merge-Tree,
Bigtable: A Distributed Storage System
for Structured Data

SSTable format, briefly

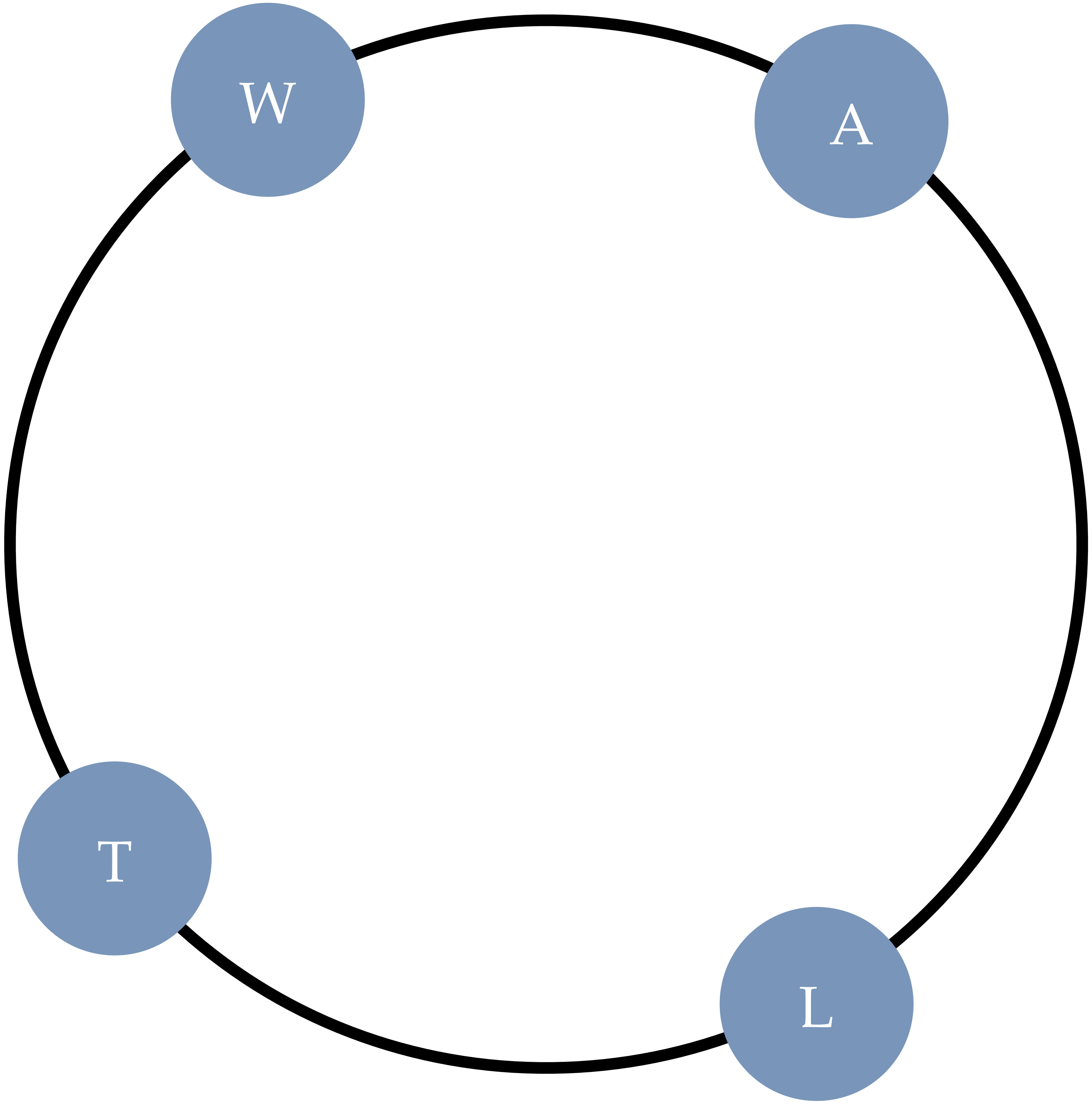


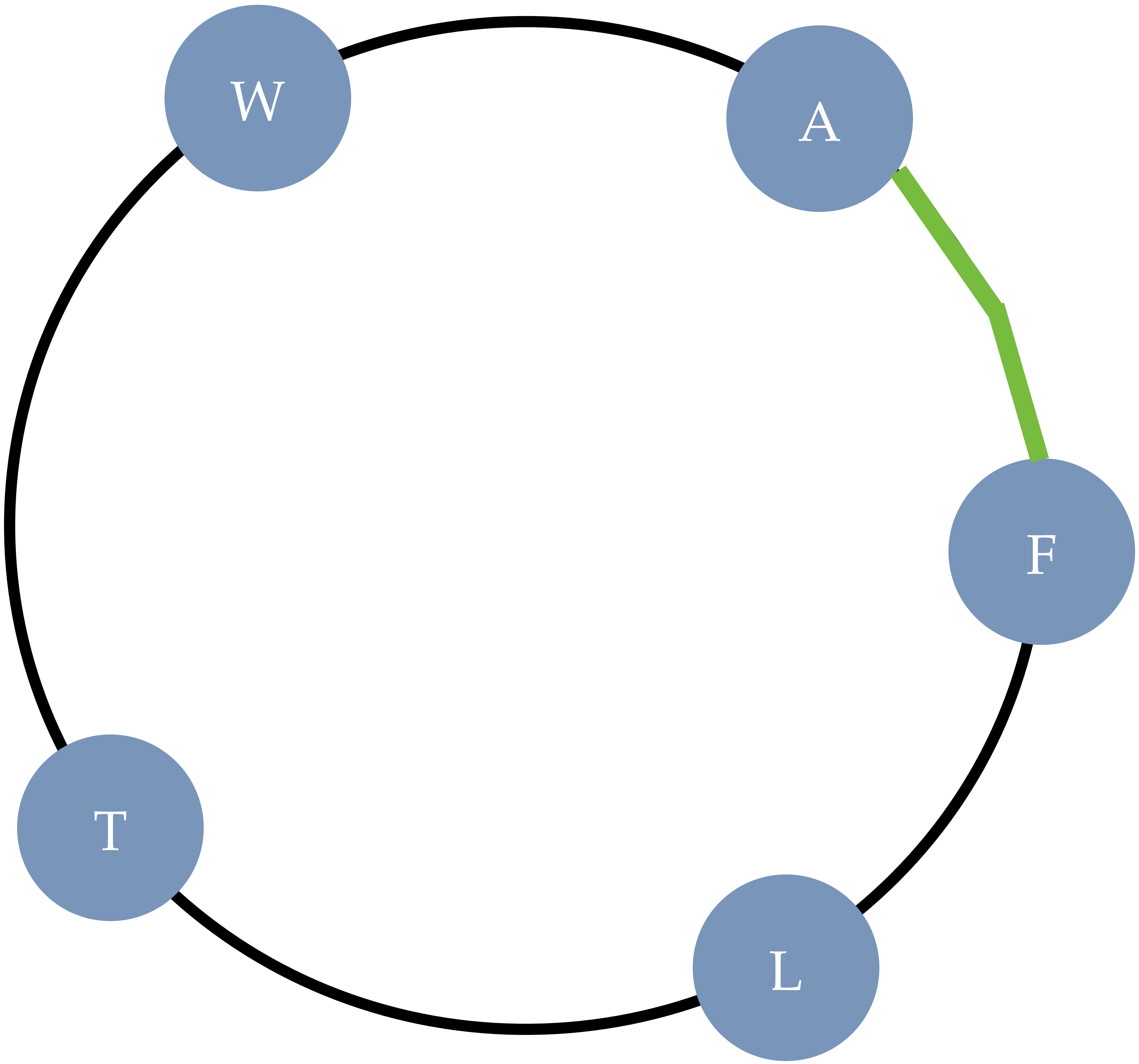
<key 127>
<key 255>
...

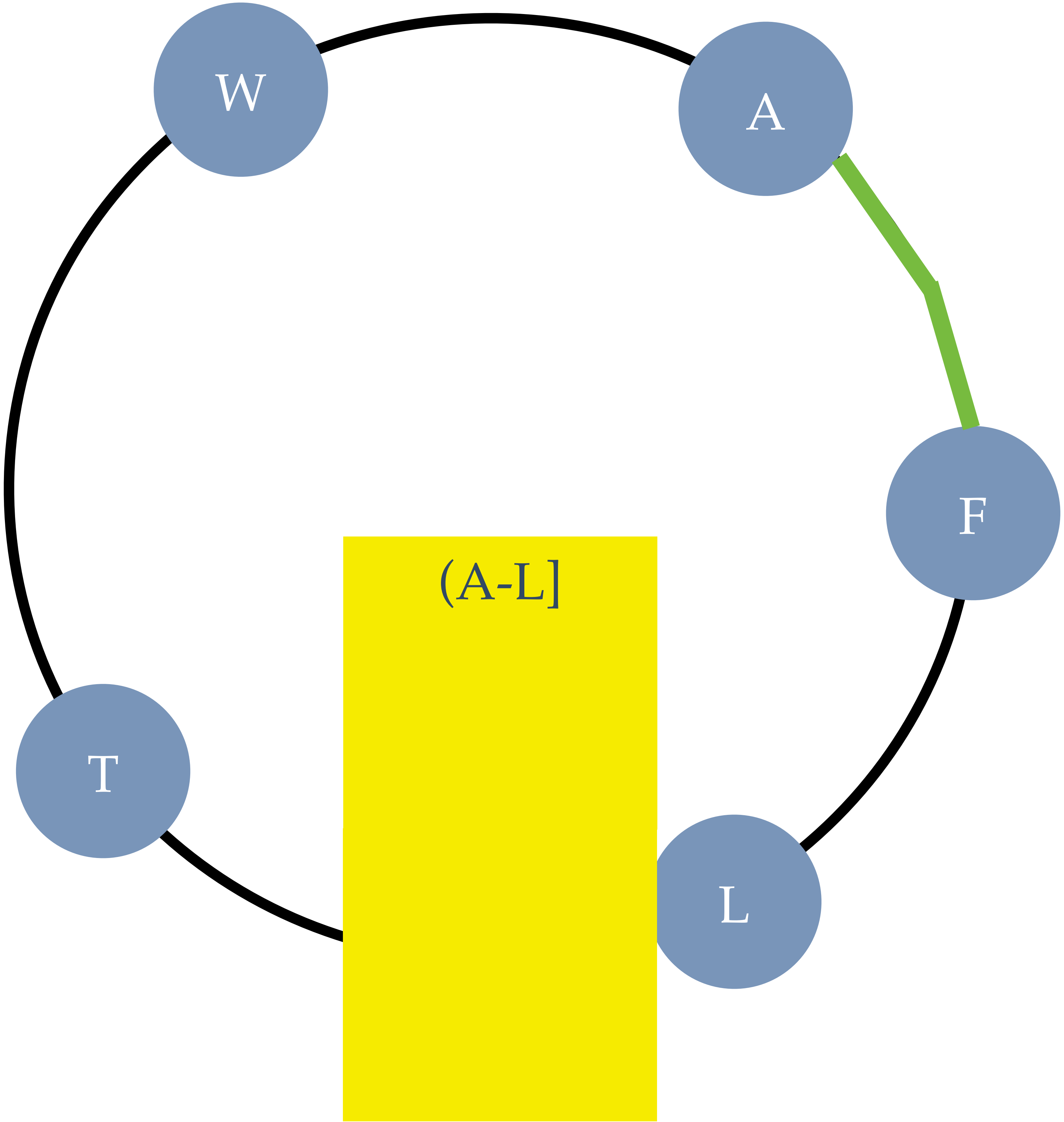
<row data 0>
<row data 1>
...
<row data 127>
...
<row data 255>
...

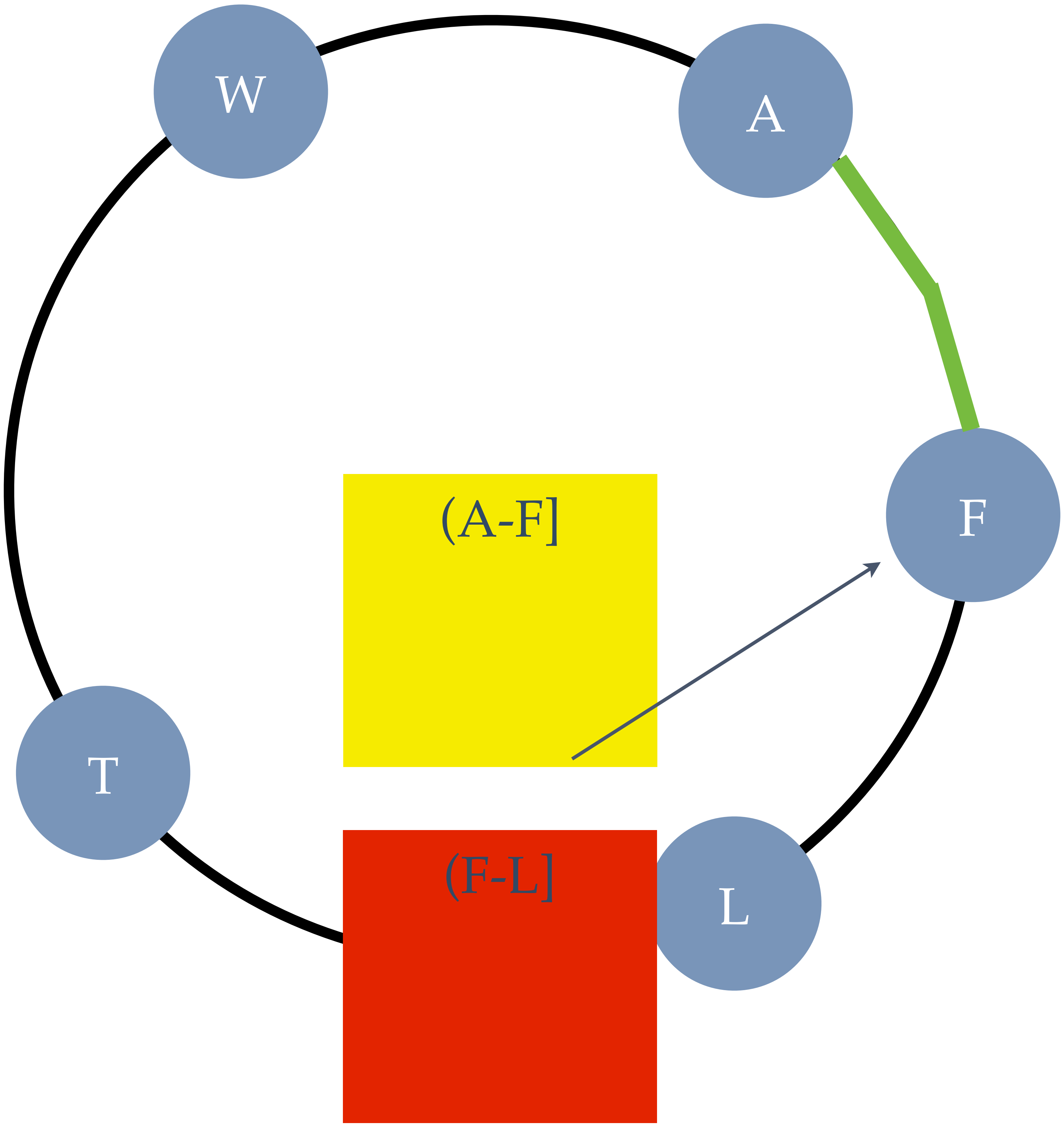
Sorted [clustered] by row key

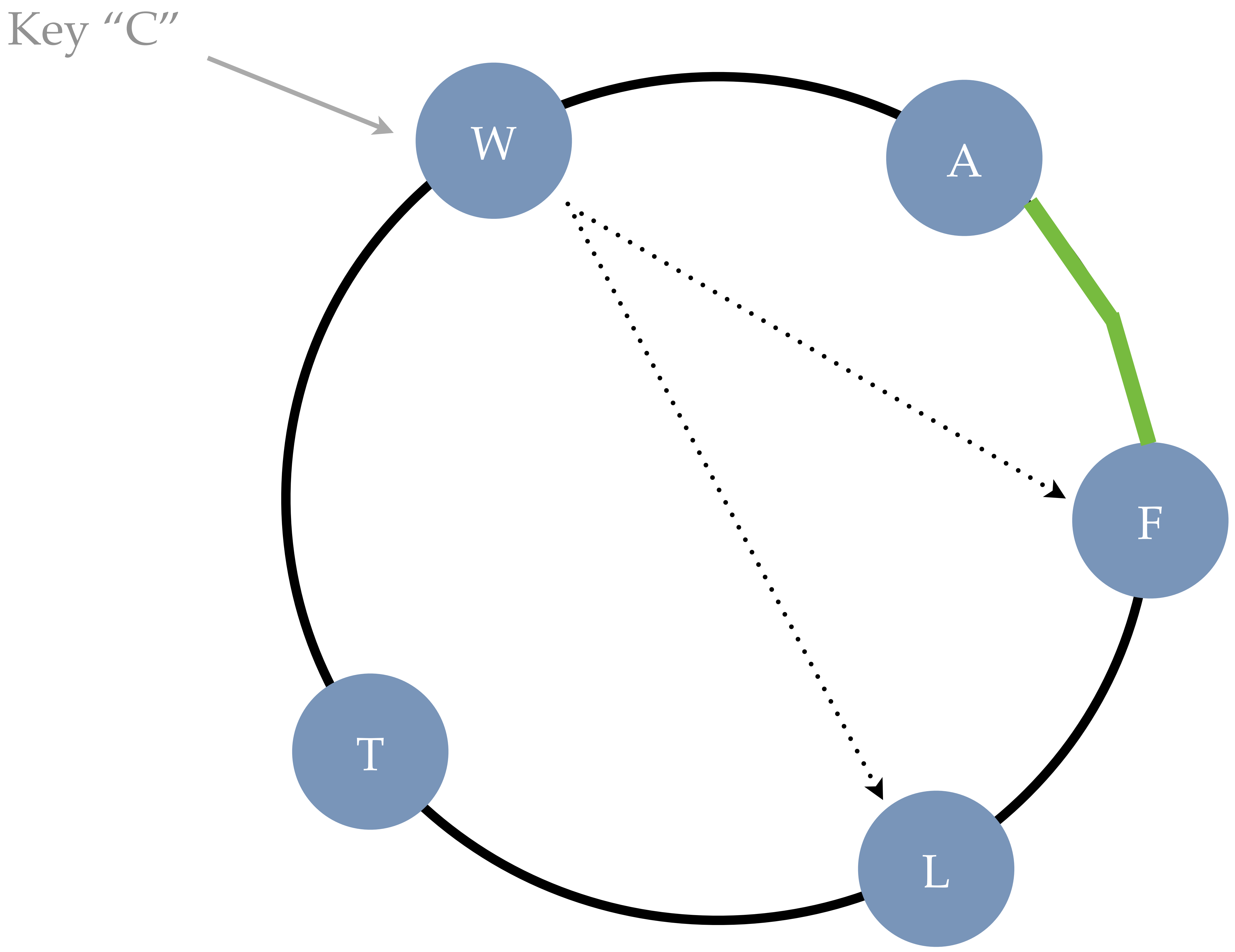
Scaling


W
A
T
L


W
A
F
T
L


W
A
F
(A-L]
T
L


W
A
F
(A-F]
T
(F-L]
L

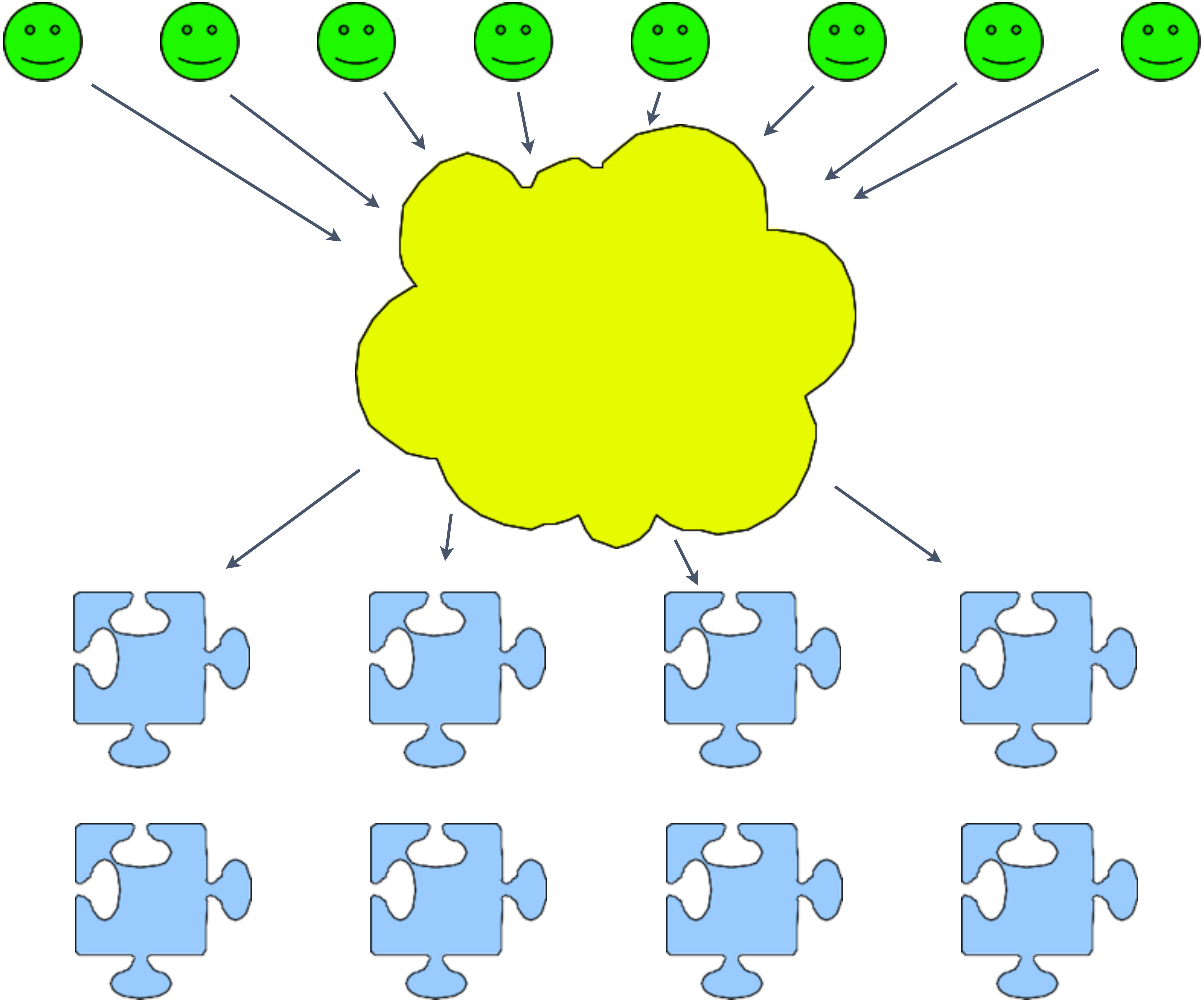

Key "C"
W
A
F
T
L

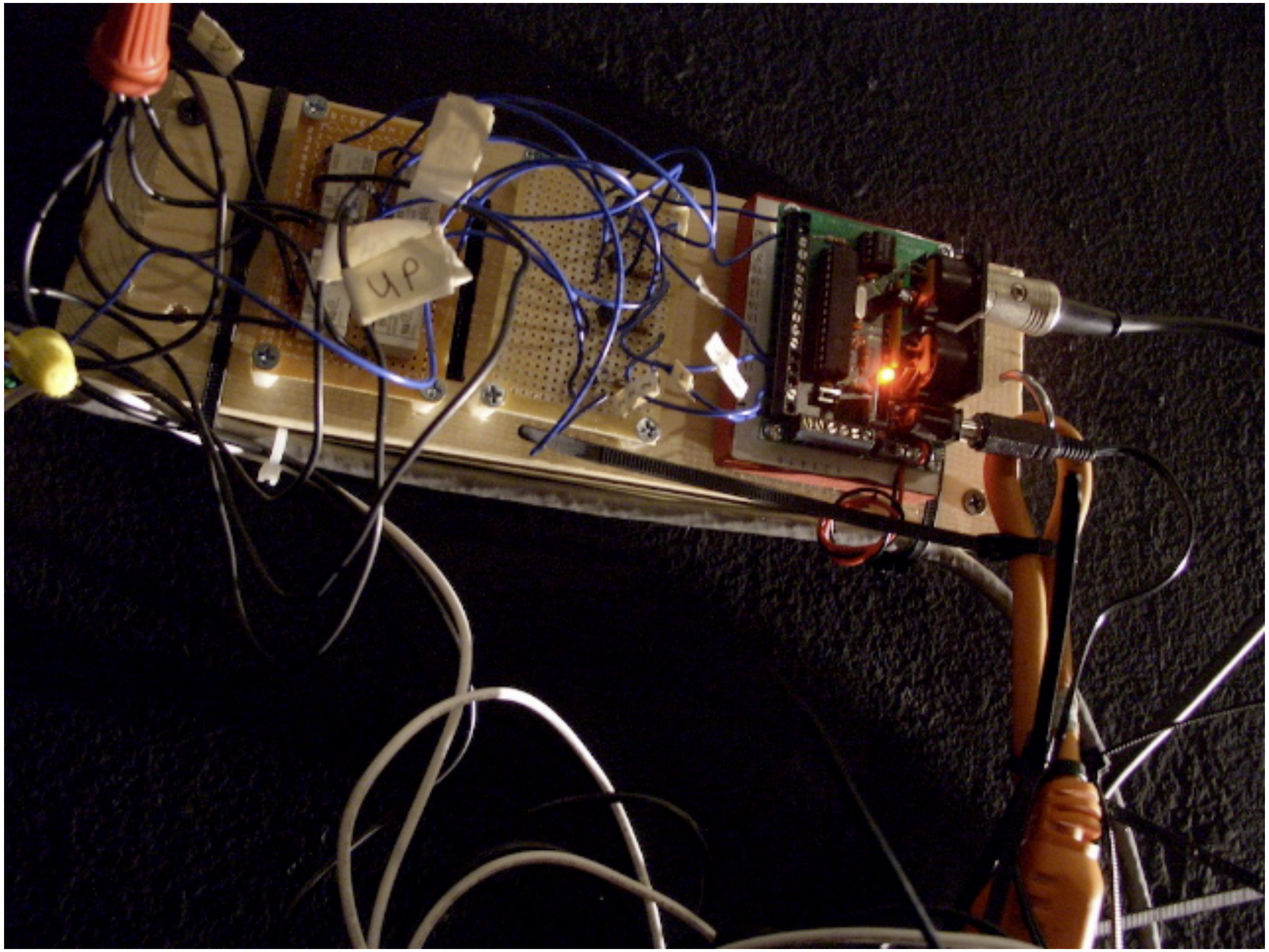
Reliability

- No single points of failure
- Multiple datacenters
- Monitorable

Some headlines

- “Resyncing Broken MySQL Replication”
- “How To Repair MySQL Replication”
- “Fixing Broken MySQL Database Replication”
- “Replication on Linux broken after db restore”
- “MySQL :: Repairing broken replication”



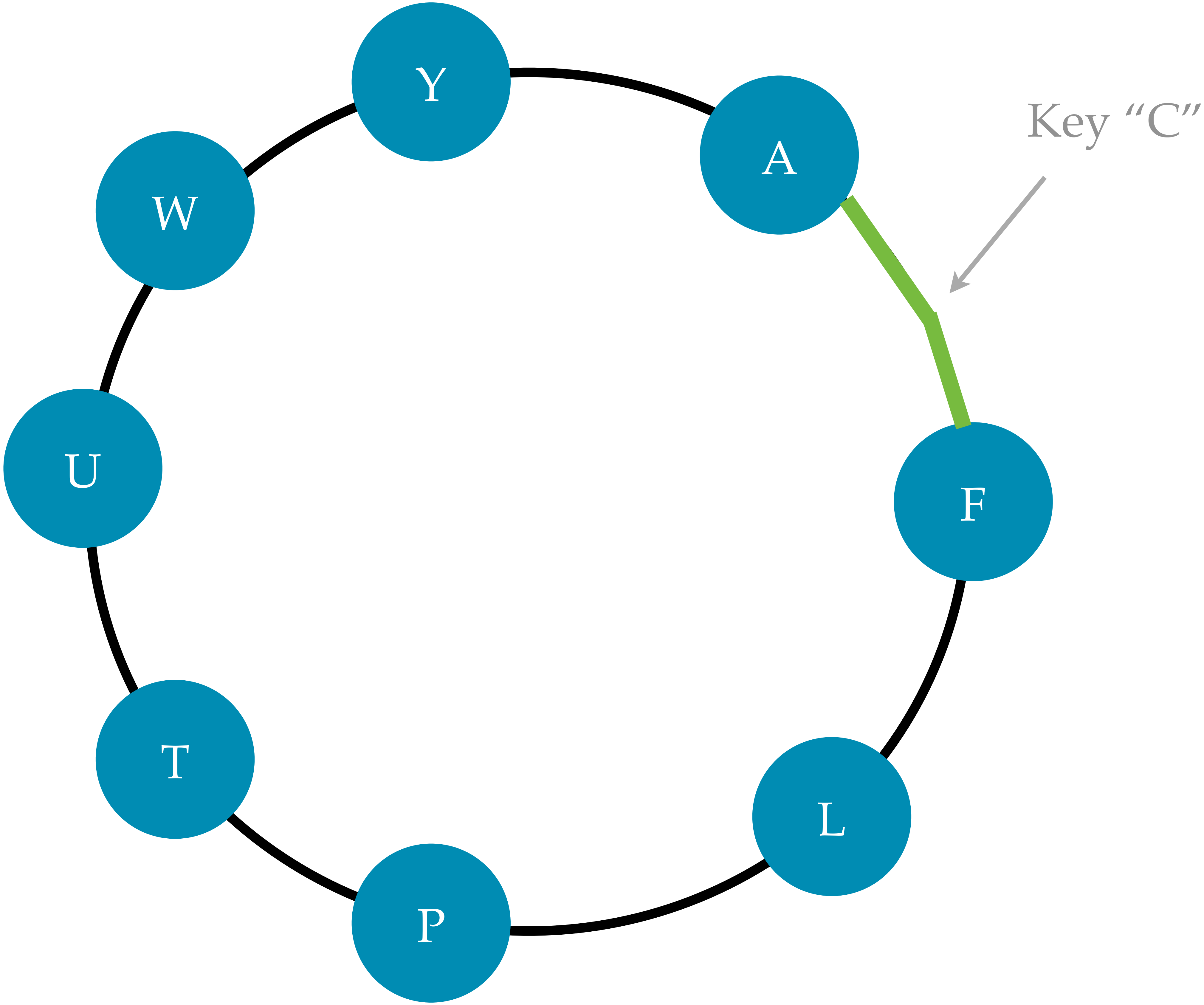

up

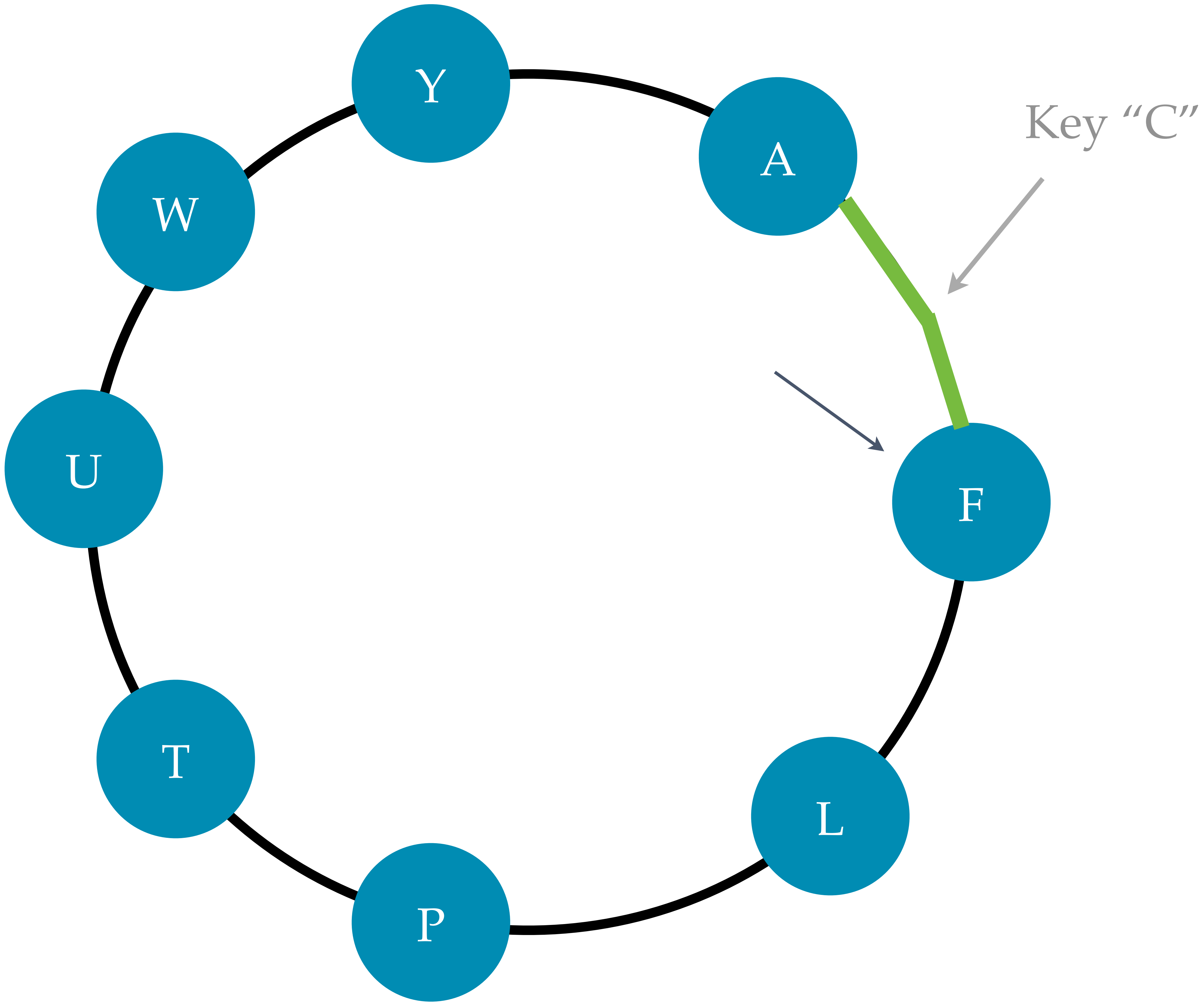
The opposite of heroes

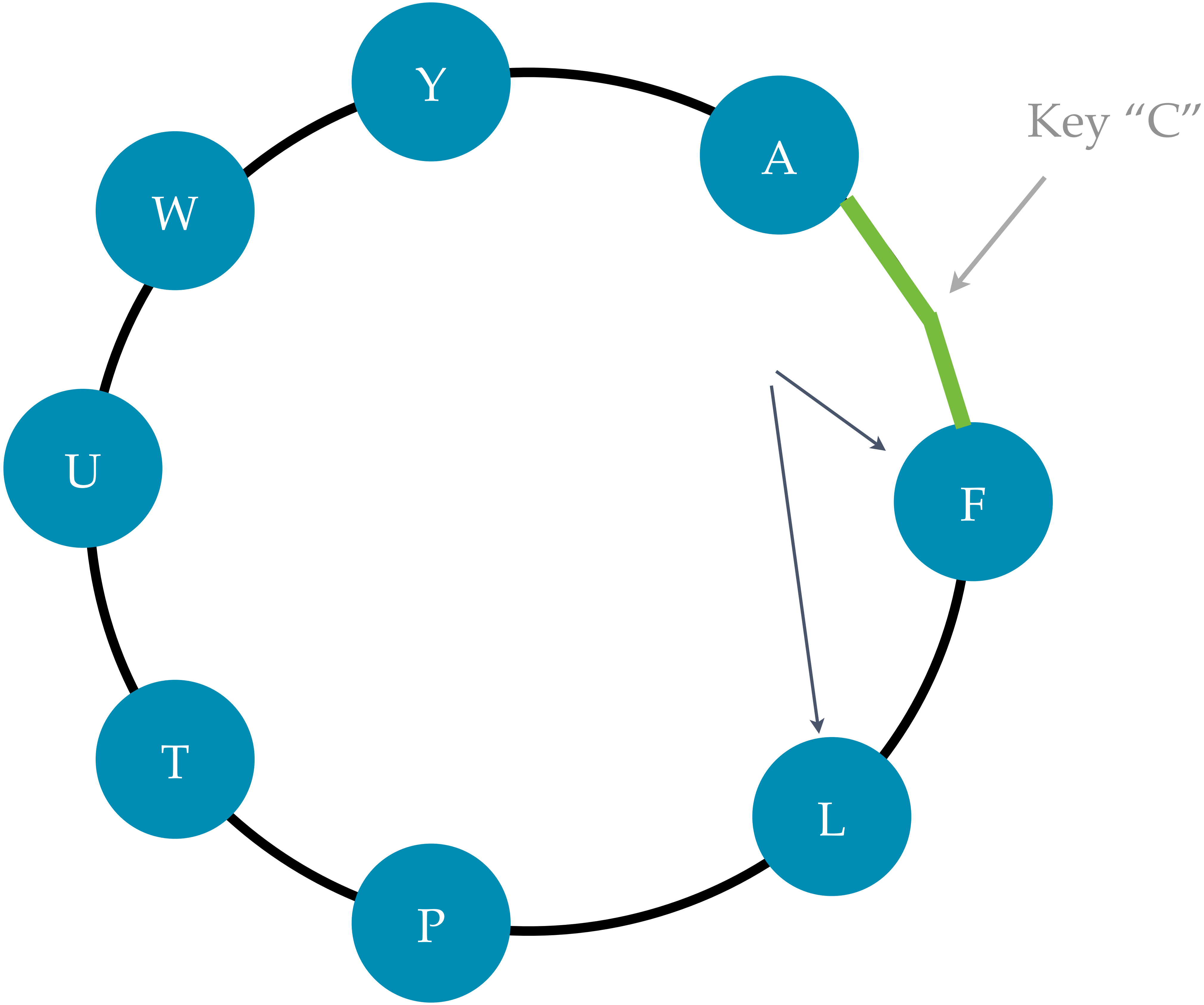
- “If your software wakes someone up at 4 AM to fix it, you’re doing it wrong.”

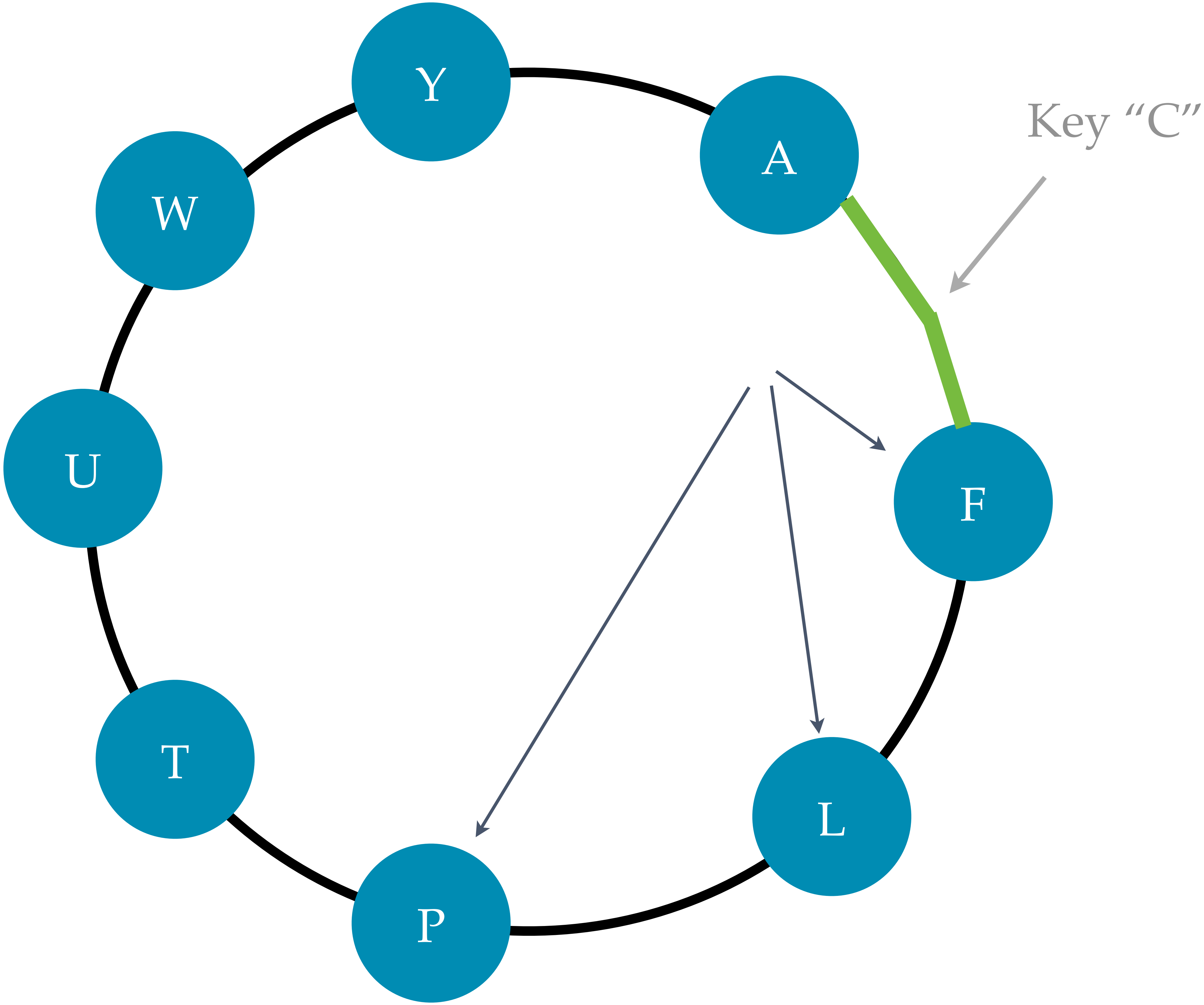
Good architecture solves multiple problems at once

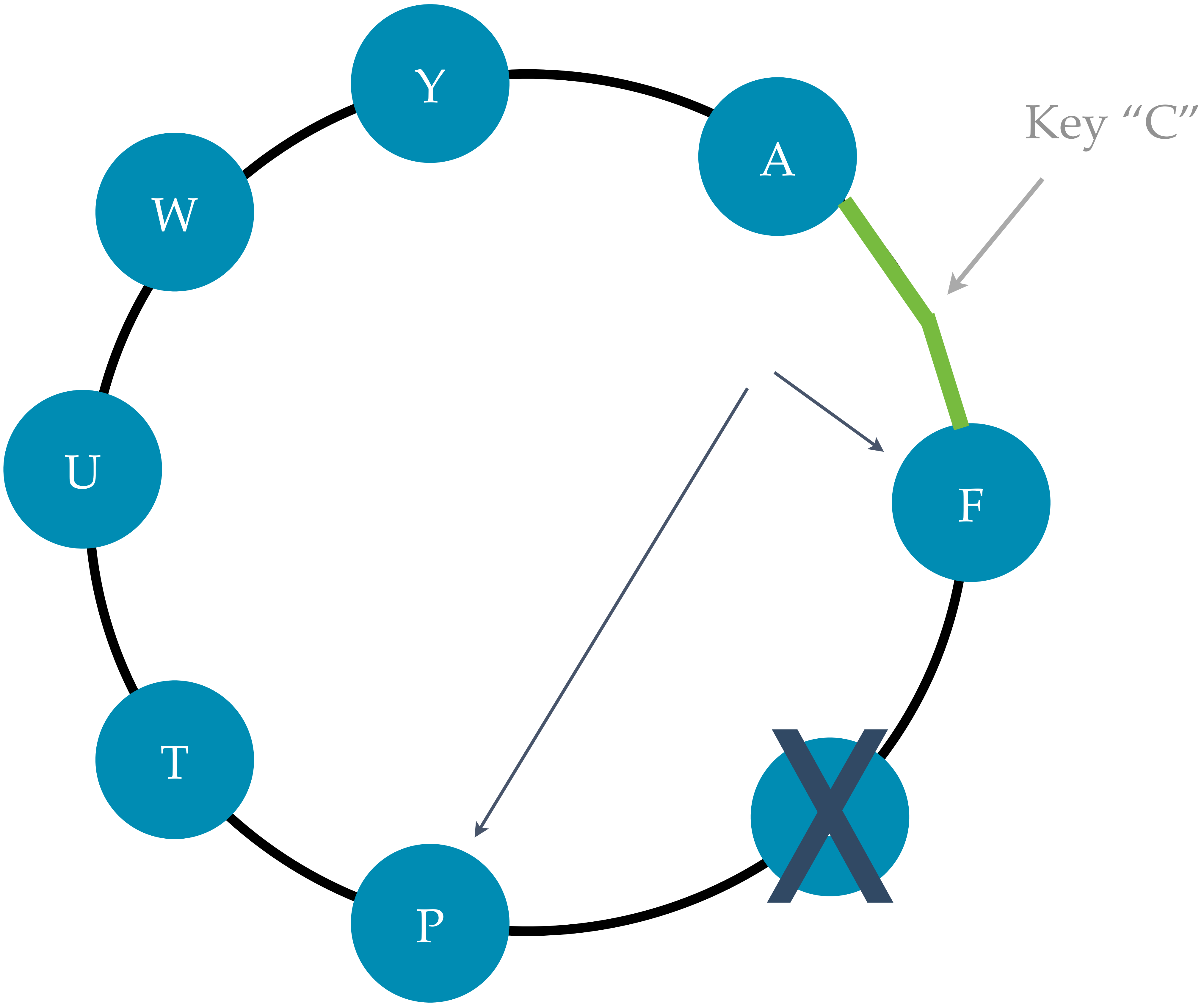
- Availability in single datacenter
- Availablility in multiple datacenters

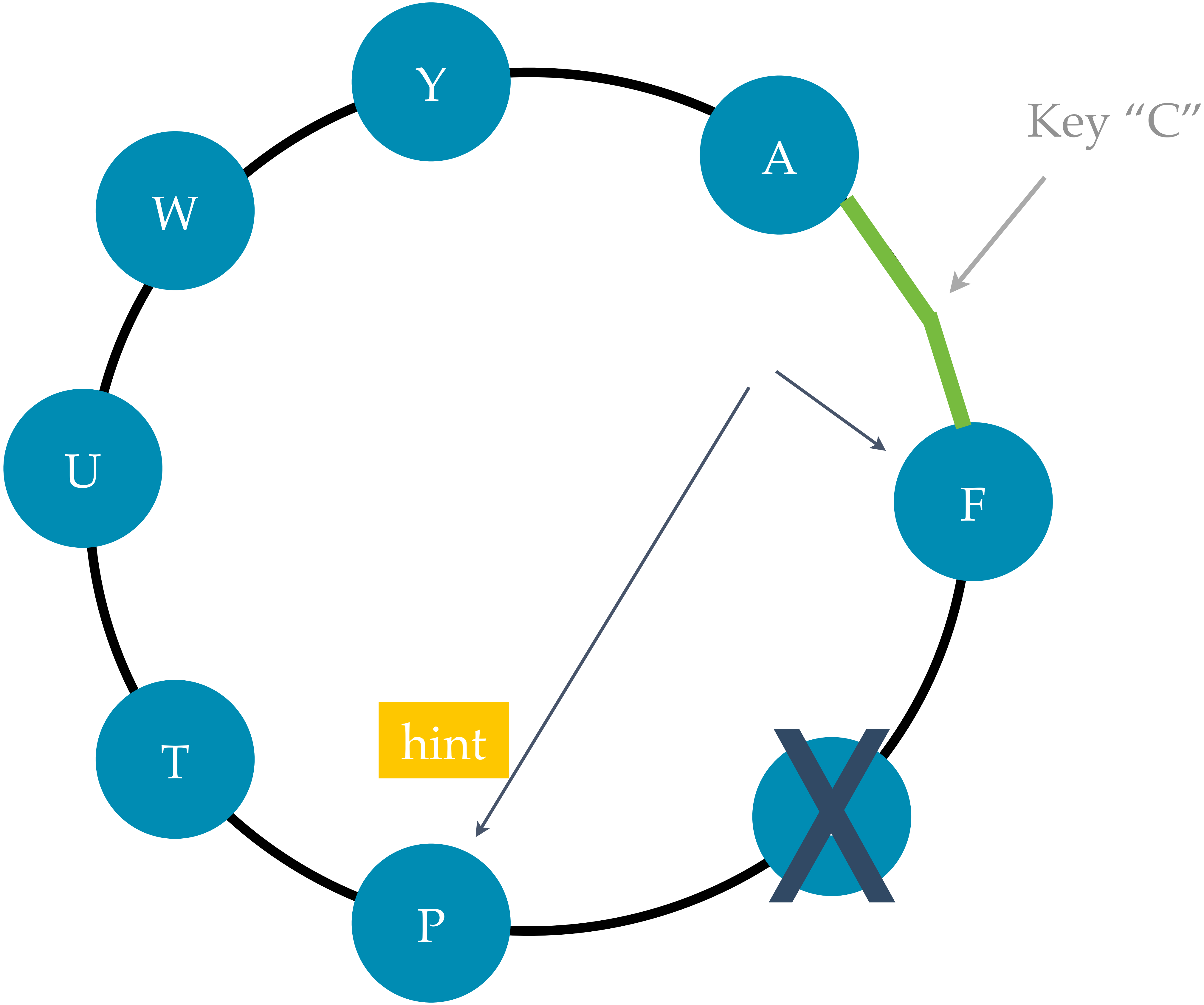

Y
A
Key "C"
W
U
F
T
L
P

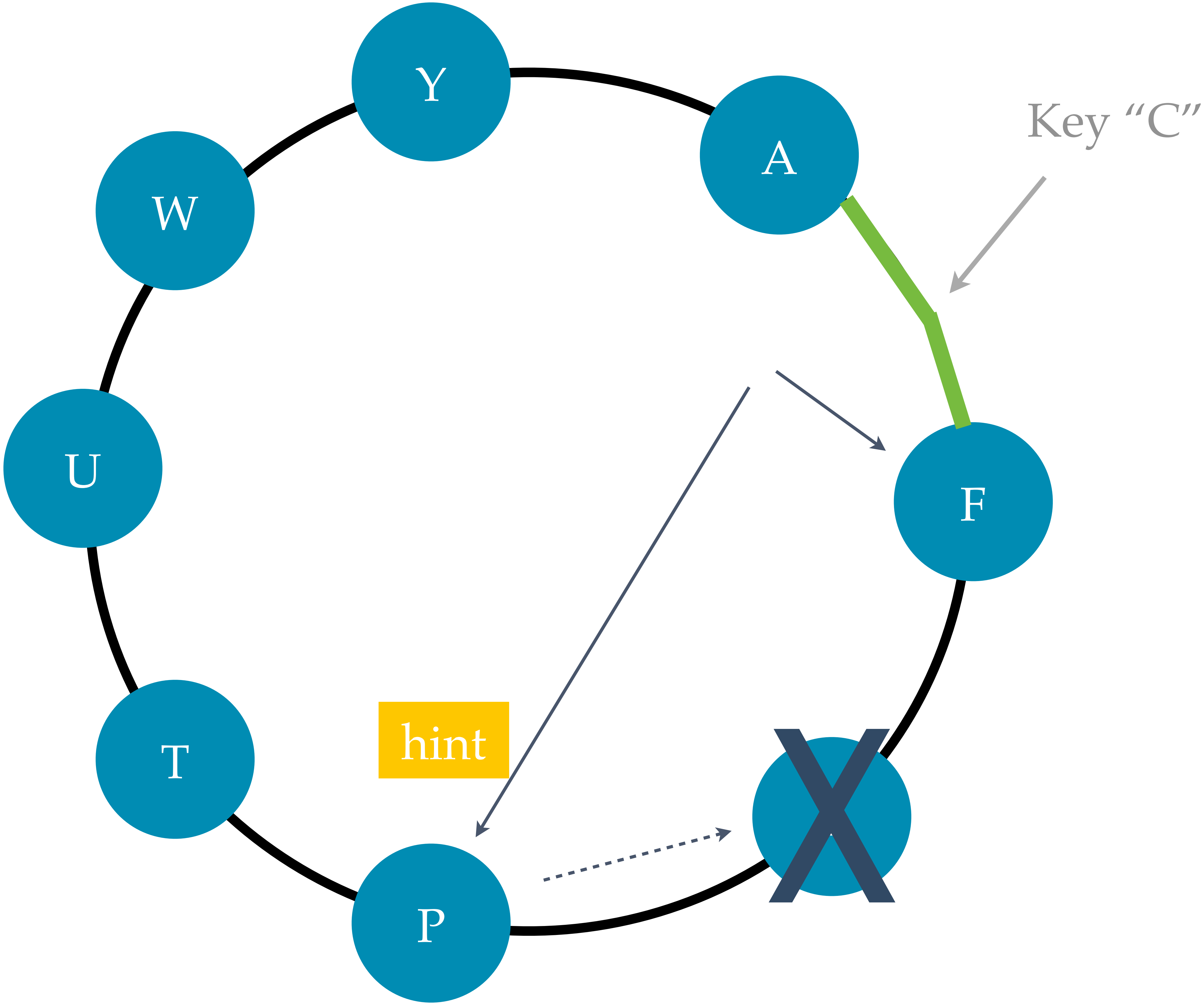

Y
A
Key "C"
W
U
F
T
L
P

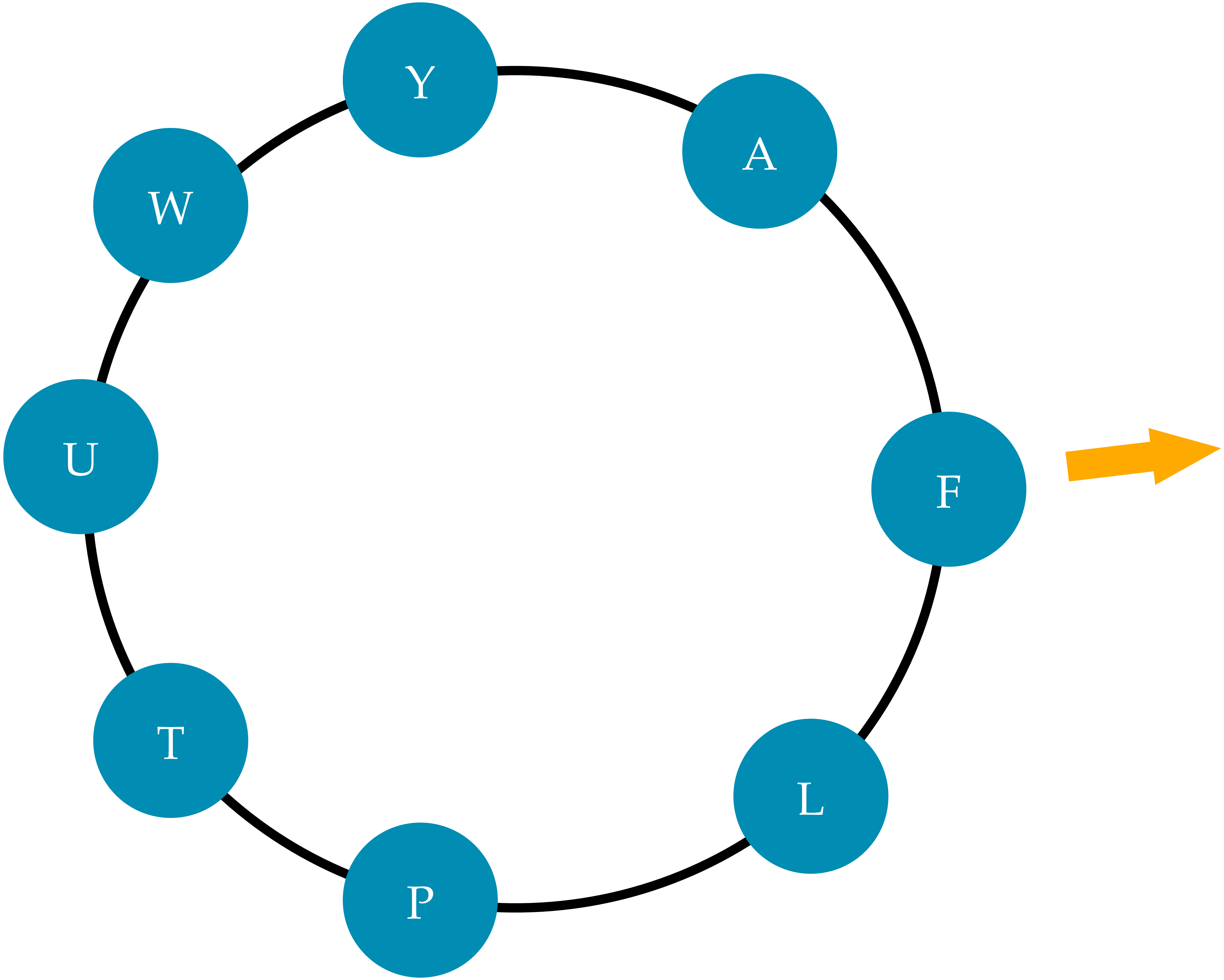

Y
A
W
U
F
T
L
P
Key "C"

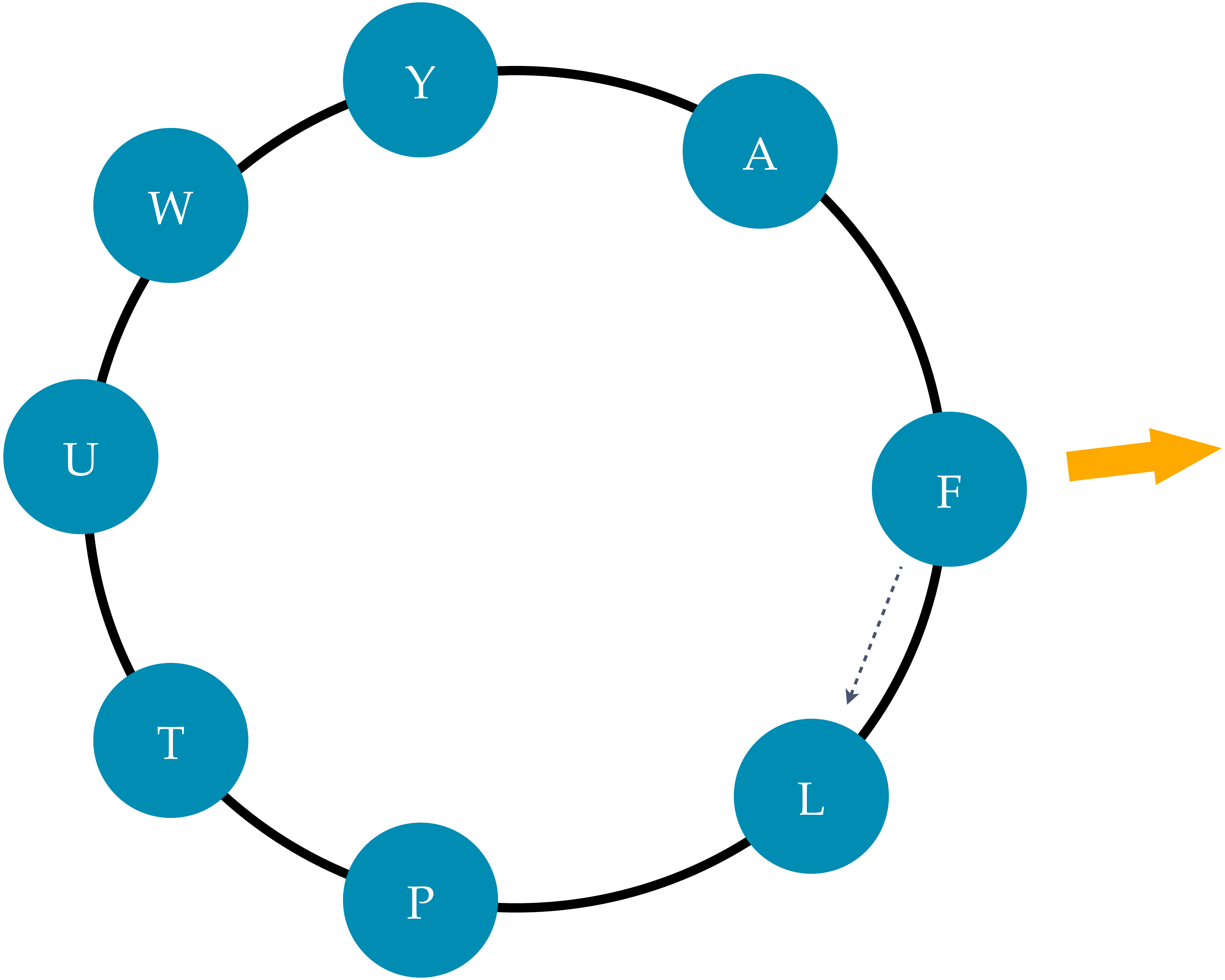

Y
A
Key "C"
W
U
F
T
L
P

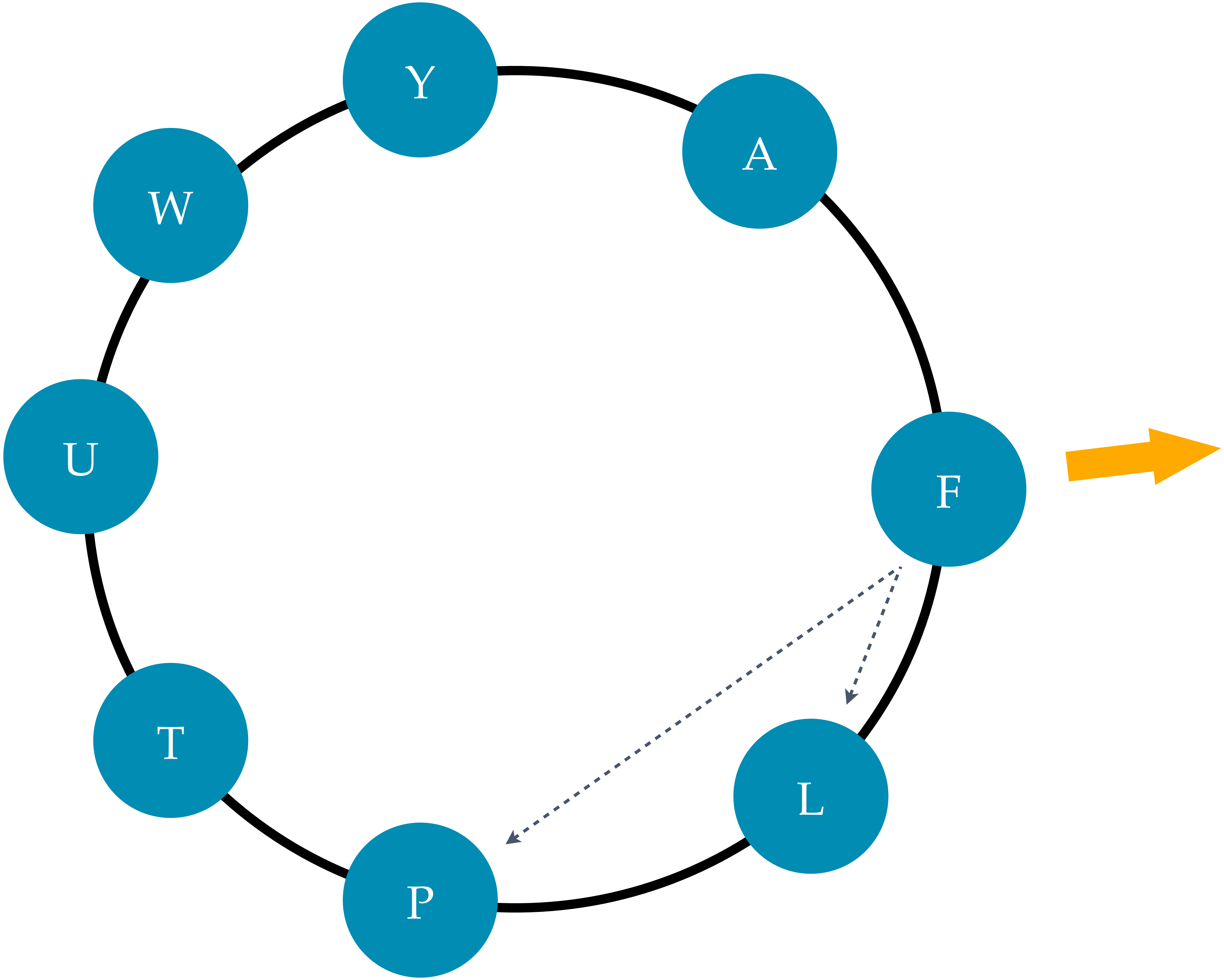

Y
A
Key "C"
W
U
F
T
X
P

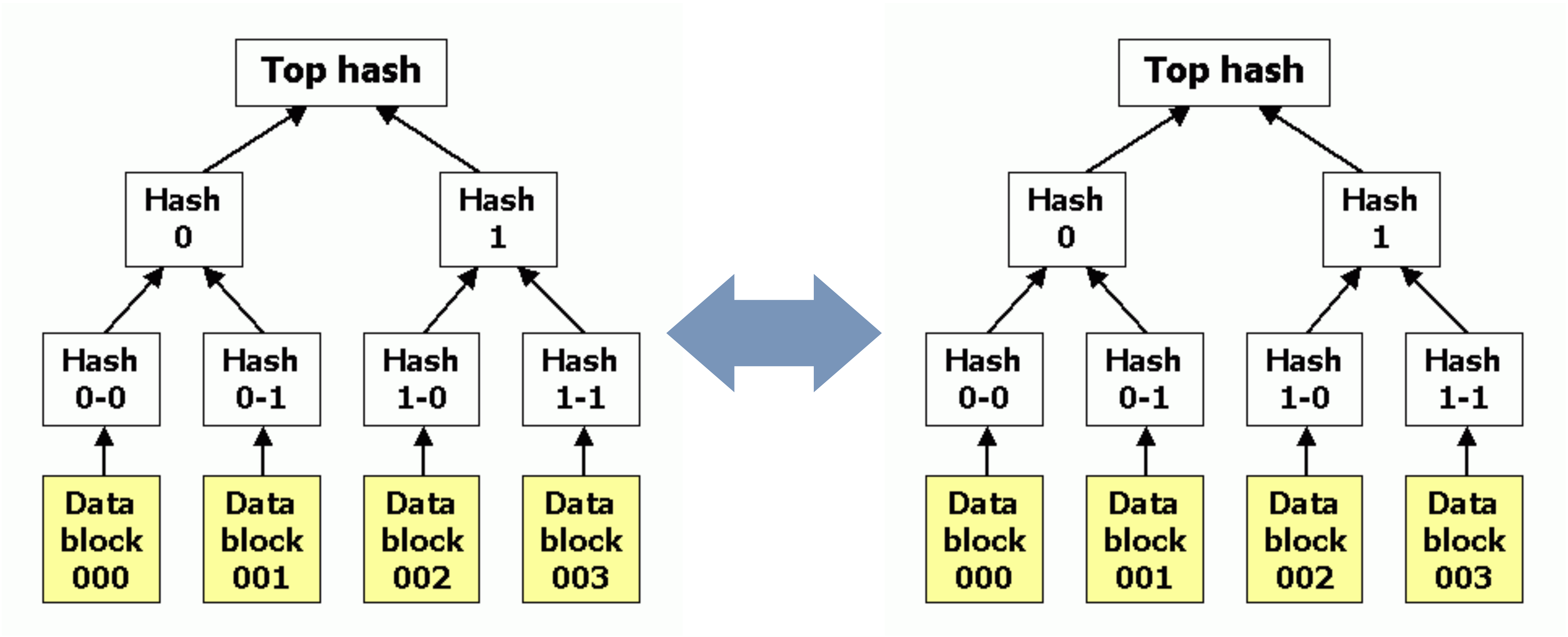

Y
A
Key “C”
W
U
F
T
hint
X
P

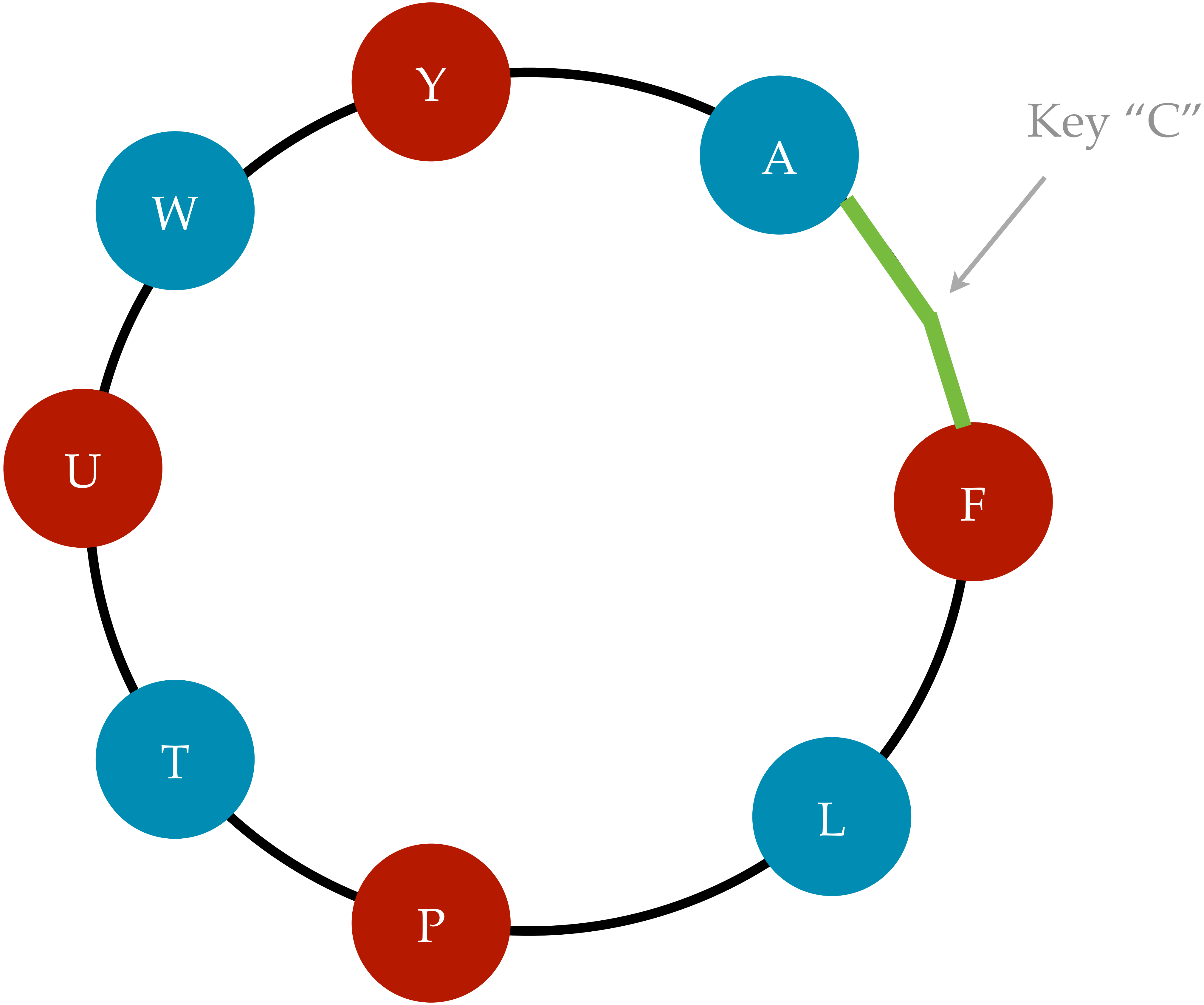

Y
A
Key "C"
W
U
F
hint
T
X
P

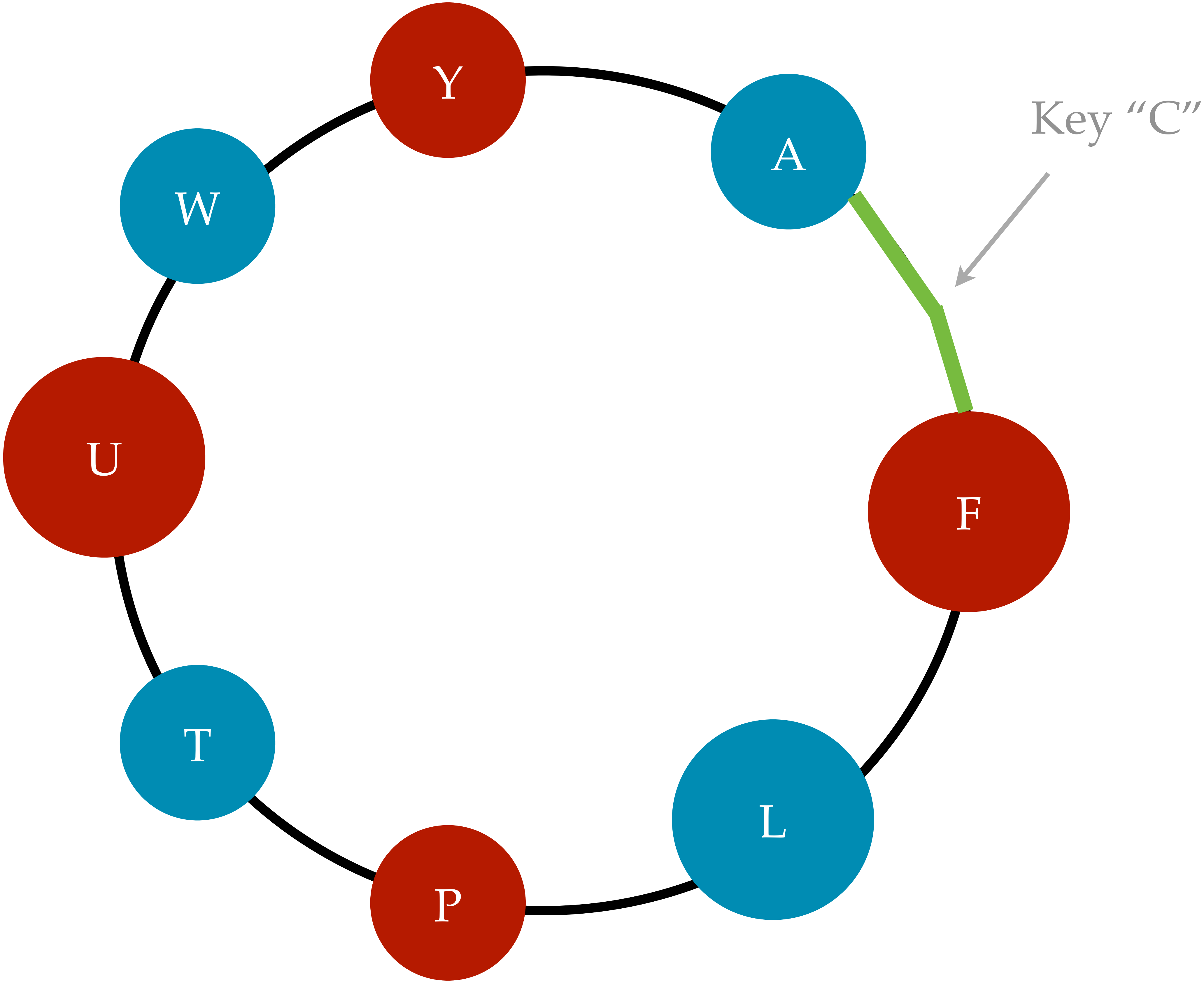

Y
A
W
U
F
T
L
P


Y
A
W
U
F
T
L
P


Y
A
W
U
F
T
L
P


Top hash
Hash 0
Hash 1
Hash 0-0
Hash 0-1
Hash 1-0
Hash 1-1
Data block 000
Data block 001
Data block 002
Data block 003
Top hash
Hash 0
Hash 1
Hash 0-0
Hash 0-1
Hash 1-0
Hash 1-1
Data block 000
Data block 001
Data block 002
Data block 003

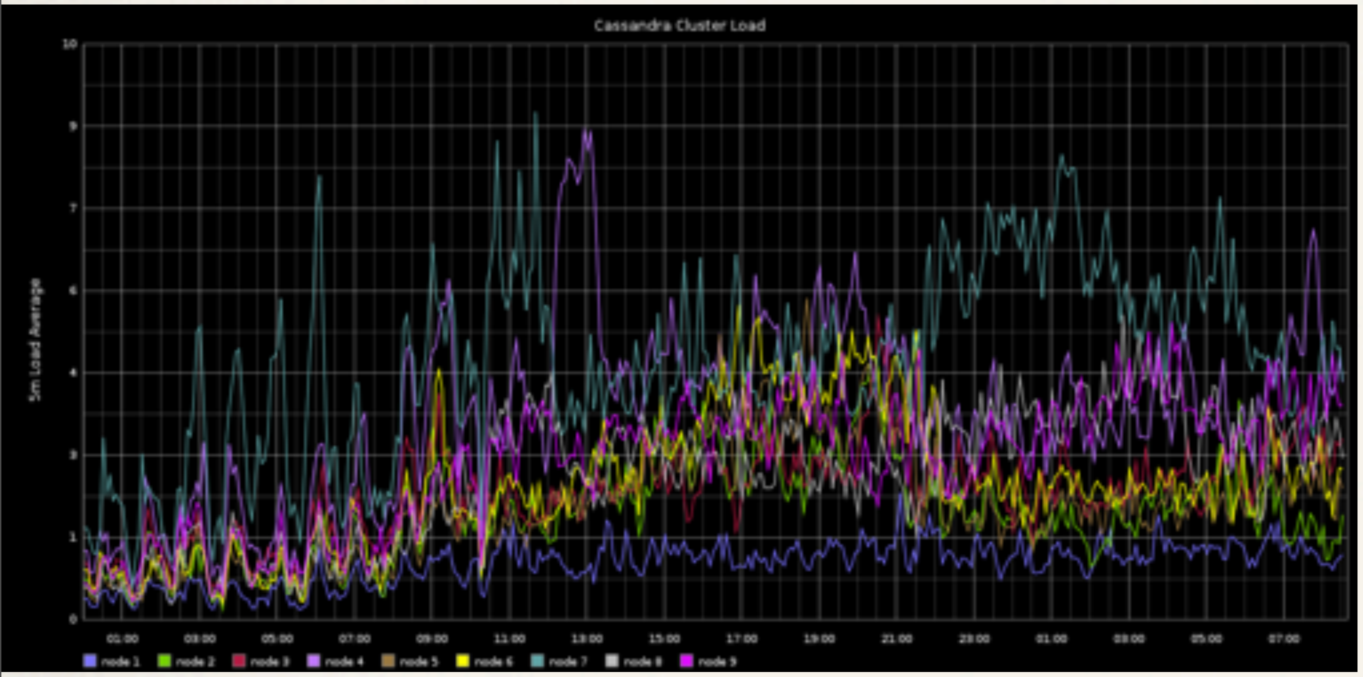

Y
A
Key "C"
W
U
F
T
L
P

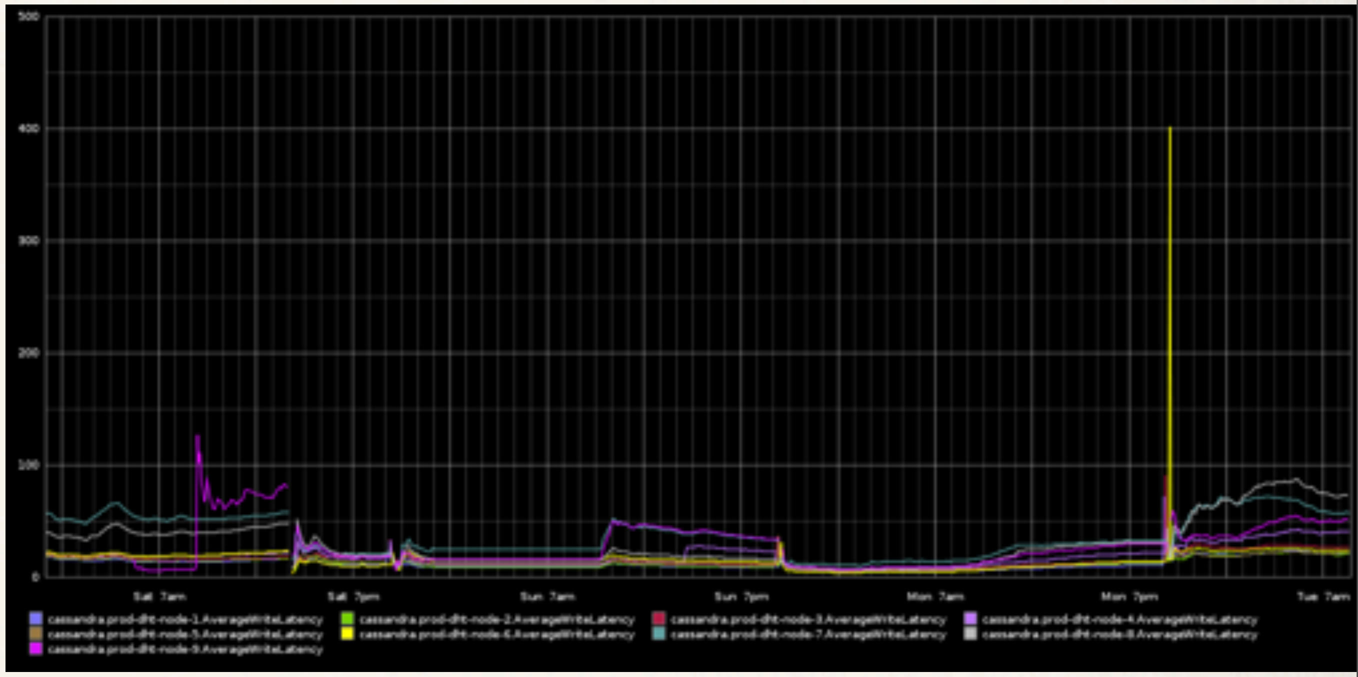

Y
A
Key "C"
W
U
F
T
L
P

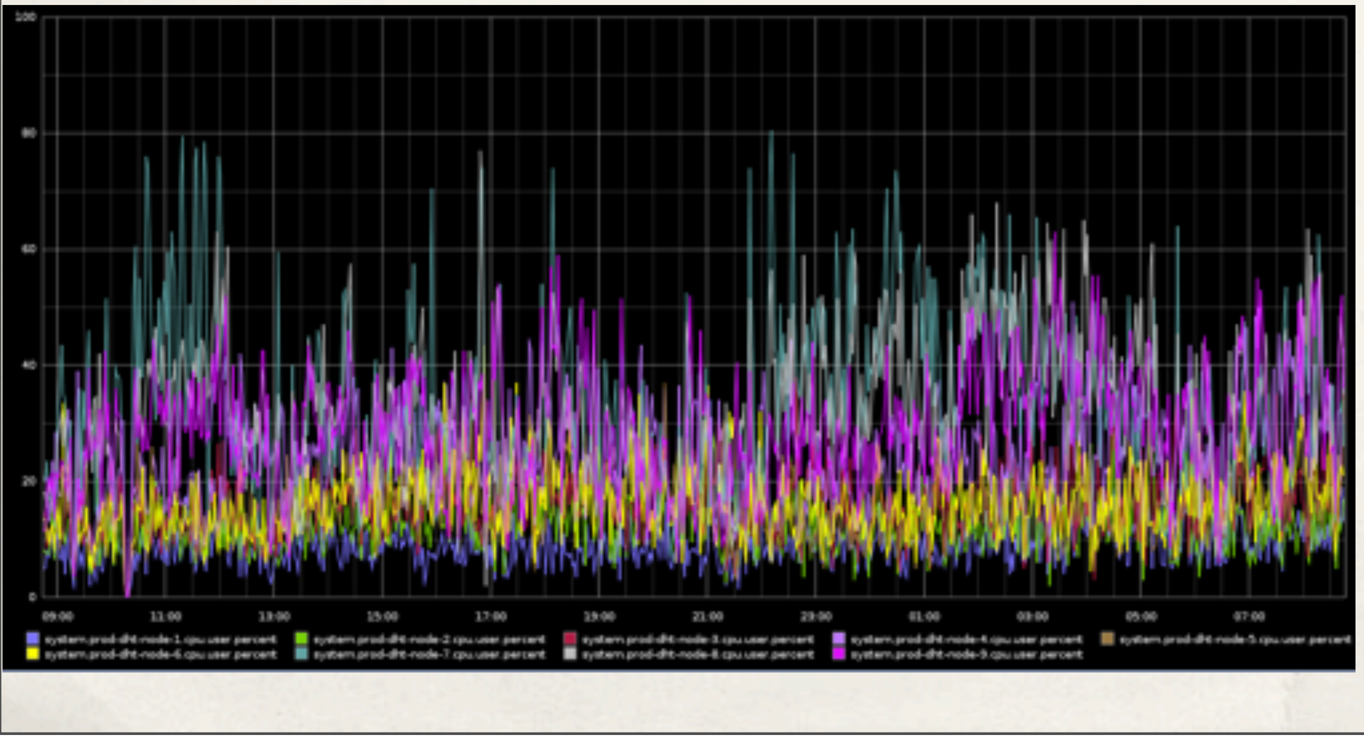
Tuneable consistency

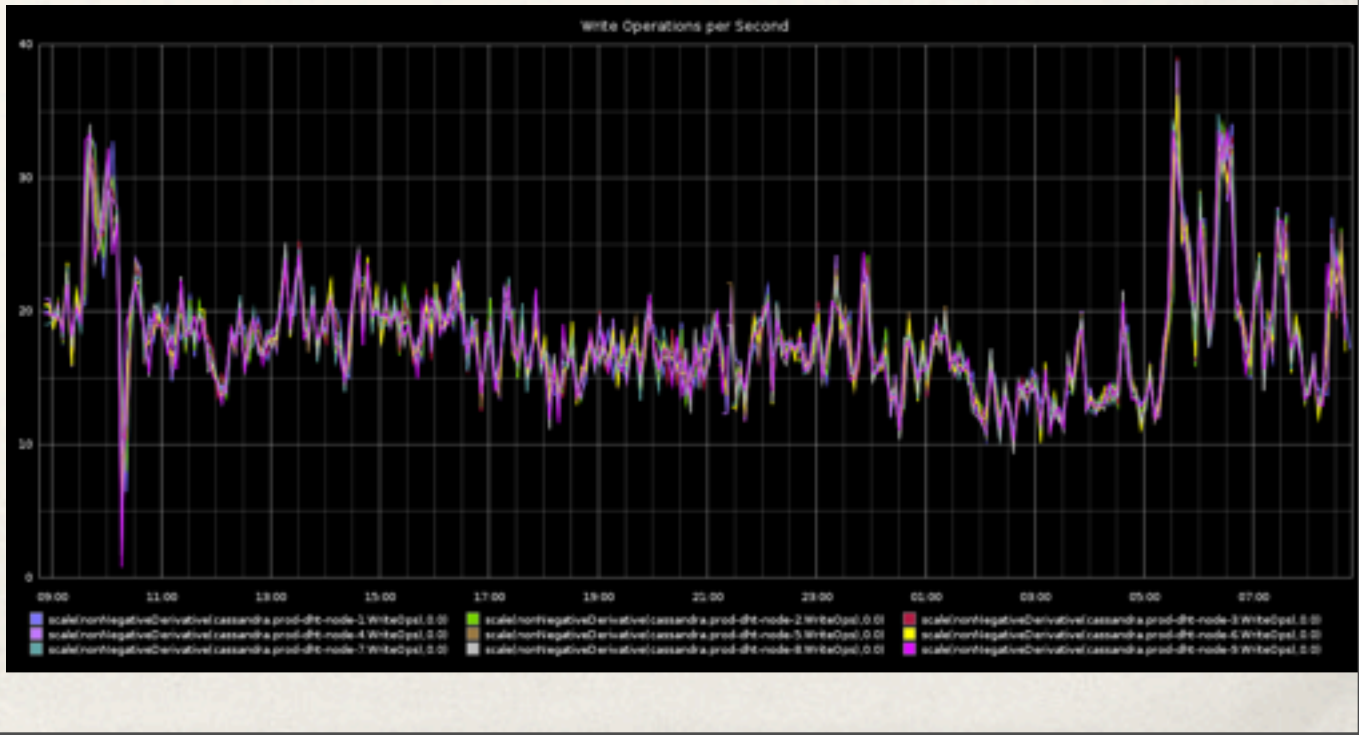
- ONE, QUORUM, ALL
- R + W > N
- Choose availability vs consistency (and latency)

Monitorable

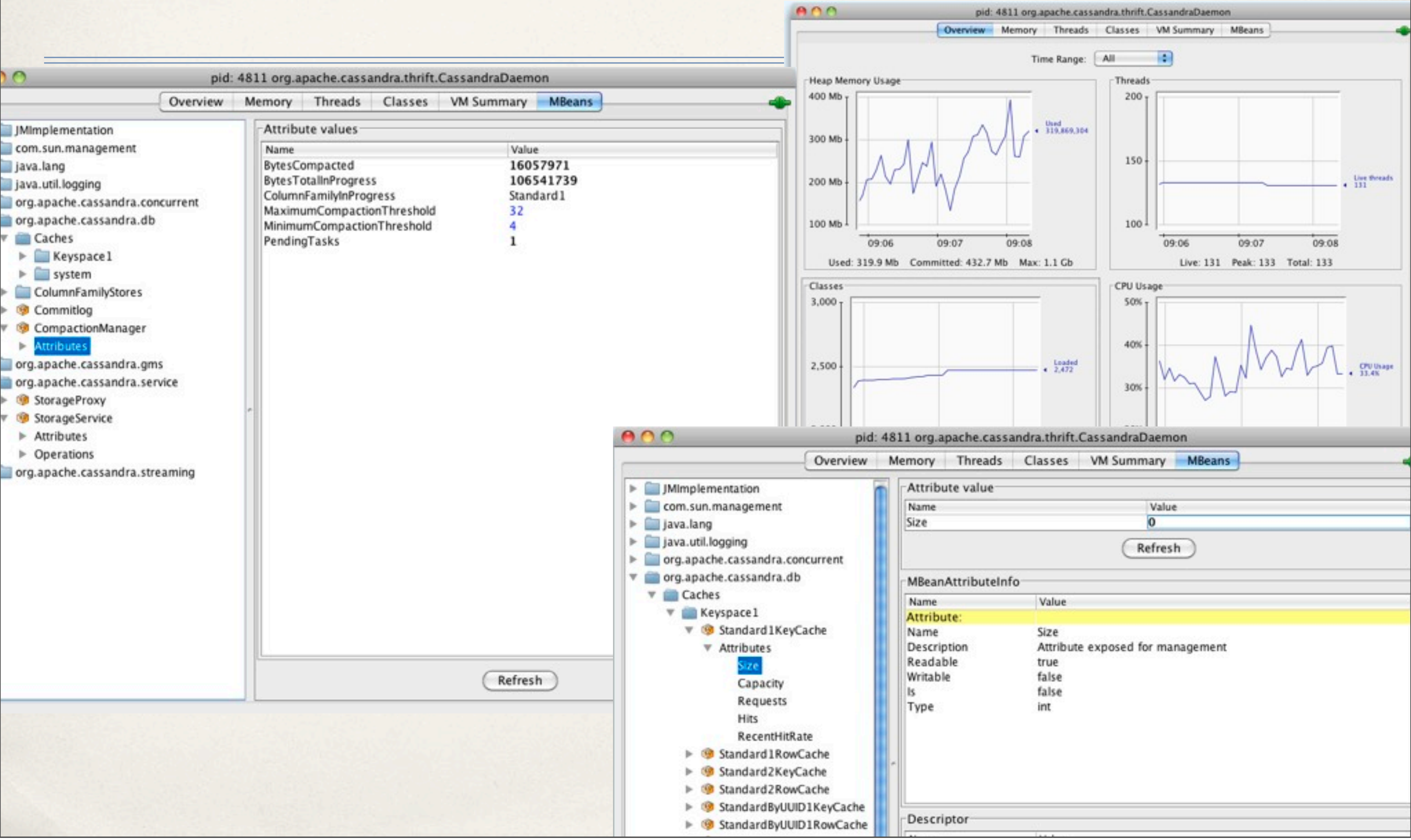








JMX



Used: 319.9 Mb Committed: 432.7 Mb Max: 1.1 Gb

Live: 131 Peak: 133 Total: 133

Attribute values

Name	Value
BytesCompacted	**16057971**
BytesTotalInProgress	**106541739**
ColumnFamilyInProgress	Standard1
MaximumCompactionThreshold	32
MinimumCompactionThreshold	4
PendingTasks	**1**

Attribute value

Name	Value
Size	0

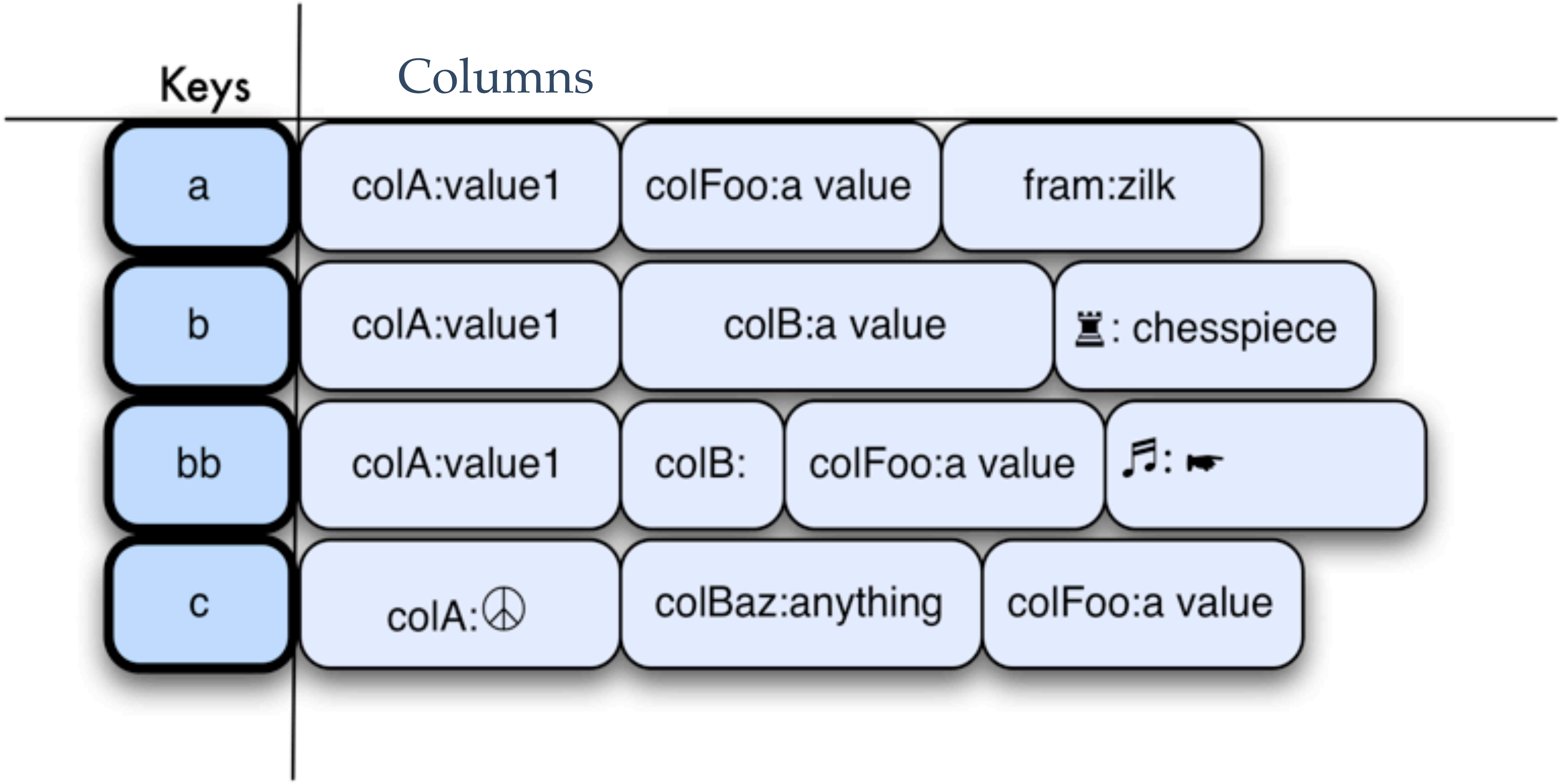
MBeanAttributeInfo

Name	Value
Attribute:	
Name	Size
Description	Attribute exposed for management
Readable	true
Writable	false
Is	false
Type	int

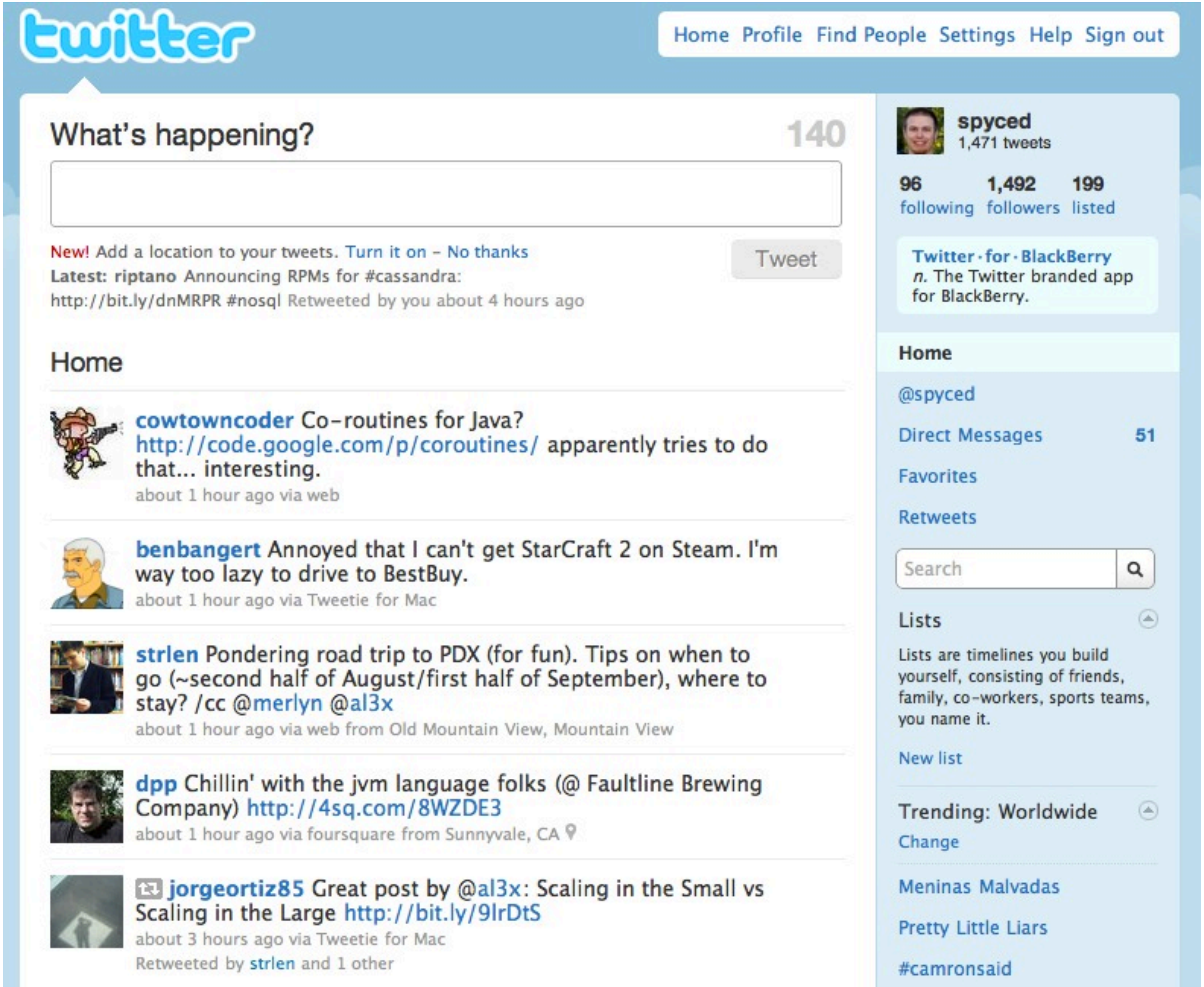
Data model tradeoffs

- Twitter: “Fifteen months ago, it took two weeks to perform ALTER TABLE on the statuses [tweets] table.”

Cassandra is not a key/value store



Keys	Columns			
a	colA:value1	colFoo:a value	fram:zilk	
b	colA:value1	colB:a value	♜: chesspiece	
bb	colA:value1	colB:	colFoo:a value	♬: ☛
c	colA:☮	colBaz:anything	colFoo:a value	



What's happening?

New! Add a location to your tweets. Turn it on – No thanks

Latest: riptano Announcing RPMs for #cassandra: http://bit.ly/dnMRPR #nosql Retweeted by you about 4 hours ago

Home

cowtowncoder Co-routines for Java? http://code.google.com/p/coroutines/ apparently tries to do that... interesting.
about 1 hour ago via web

benbangert Annoyed that I can't get StarCraft 2 on Steam. I'm way too lazy to drive to BestBuy.
about 1 hour ago via Tweetie for Mac

strlen Pondering road trip to PDX (for fun). Tips on when to go (~second half of August/first half of September), where to stay? /cc @merlyn @al3x
about 1 hour ago via web from Old Mountain View, Mountain View

dpp Chillin' with the jvm language folks (@ Faultline Brewing Company) http://4sq.com/8WZDE3
about 1 hour ago via foursquare from Sunnyvale, CA

jorgeortiz85 Great post by @al3x: Scaling in the Small vs Scaling in the Large http://bit.ly/9lrDtS
about 3 hours ago via Tweetie for Mac
Retweeted by strlen and 1 other

spyced
1,471 tweets

96 following 1,492 followers 199 listed

Twitter·for·BlackBerry
n. The Twitter branded app for BlackBerry.

Lists

Lists are timelines you build yourself, consisting of friends, family, co-workers, sports teams, you name it.

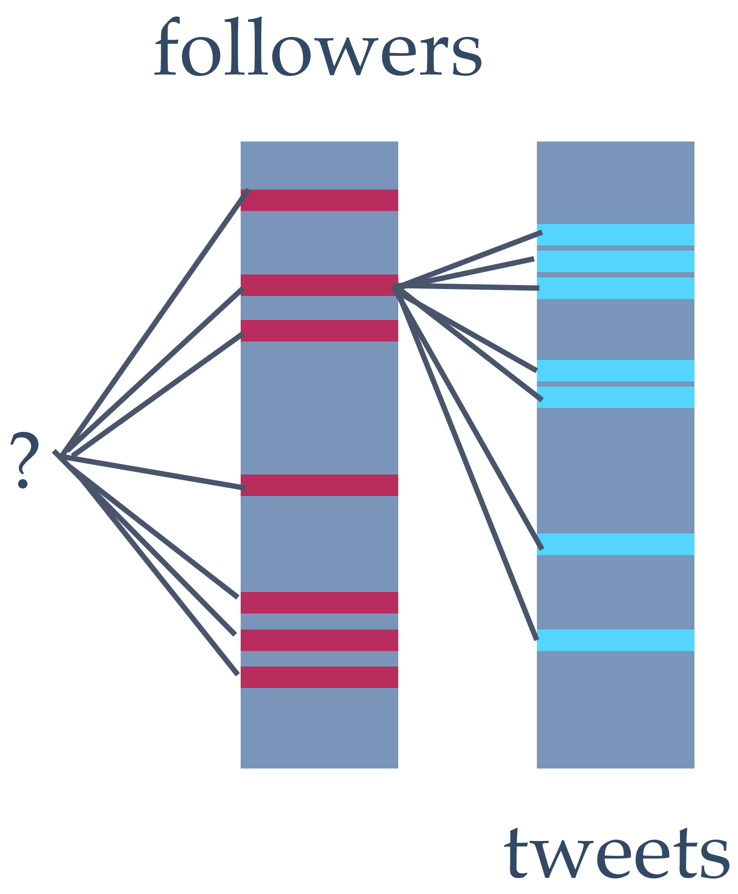
Trending: Worldwide

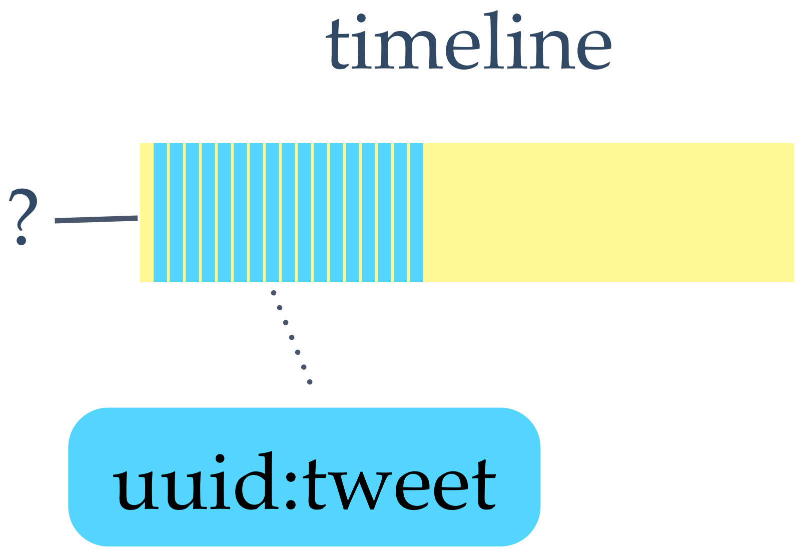
Meninas Malvadas

Pretty Little Liars

#camronsaid

```
SELECT * FROM tweets
WHERE user_id IN (SELECT follower FROM followers WHERE user_id = ?)
```





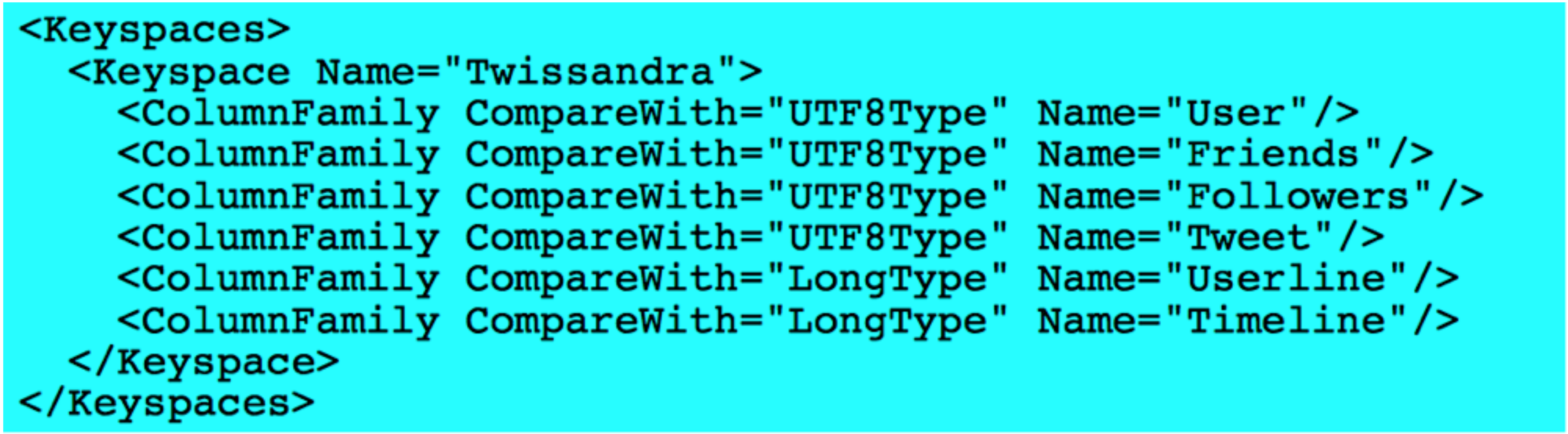
A little deeper

- http://twissandra.com
- http://github.com/jhermes/twissjava

```
CREATE TABLE users (
    id INTEGER PRIMARY KEY,
    username VARCHAR(64),
    password VARCHAR(64)
);c
```

```
CREATE TABLE following (
    user INTEGER REFERENCES user(id),
    followed INTEGER REFERENCES user(id)
);
```

```
CREATE TABLE tweets (
    id INTEGER,
    user INTEGER REFERENCES user(id),
    body VARCHAR(140),
    timestamp TIMESTAMP
);
```



```
<Keyspaces>
  <Keyspace Name="Twissandra">
    <ColumnFamily CompareWith="UTF8Type" Name="User"/>
    <ColumnFamily CompareWith="UTF8Type" Name="Friends"/>
    <ColumnFamily CompareWith="UTF8Type" Name="Followers"/>
    <ColumnFamily CompareWith="UTF8Type" Name="Tweet"/>
    <ColumnFamily CompareWith="LongType" Name="Userline"/>
    <ColumnFamily CompareWith="LongType" Name="Timeline"/>
  </Keyspace>
</Keyspaces>
```

```
Mutator m = Pelops.createMutator("Twissjava Pool",
                                 "Twissandra");

m.writeColumn(tweetid,
              TWEET,
              m.newColumn("uname", uname));
m.writeColumn(tweetid,
              TWEET,
              m.newColumn("body", body));

for (String follower : getFollowers(uname)) {
    m.writeColumn(follower,
                  TIMELINE,
                  m.newColumn(timestamp, tweetid);
}

m.execute(ConsistencyLevel.ONE);
```

```
Selector s = Pelops.createSelector("Twissjava Pool”,
                                   "Twissandra”);

s.getColumnsFromRow(uname,
                    "Timeline”,
                    s.newColumnsPredicate(startTimestamp,
                                          new byte[0],
                                          True,
                                          40),
                    ConsistencyLevel.ONE);
```

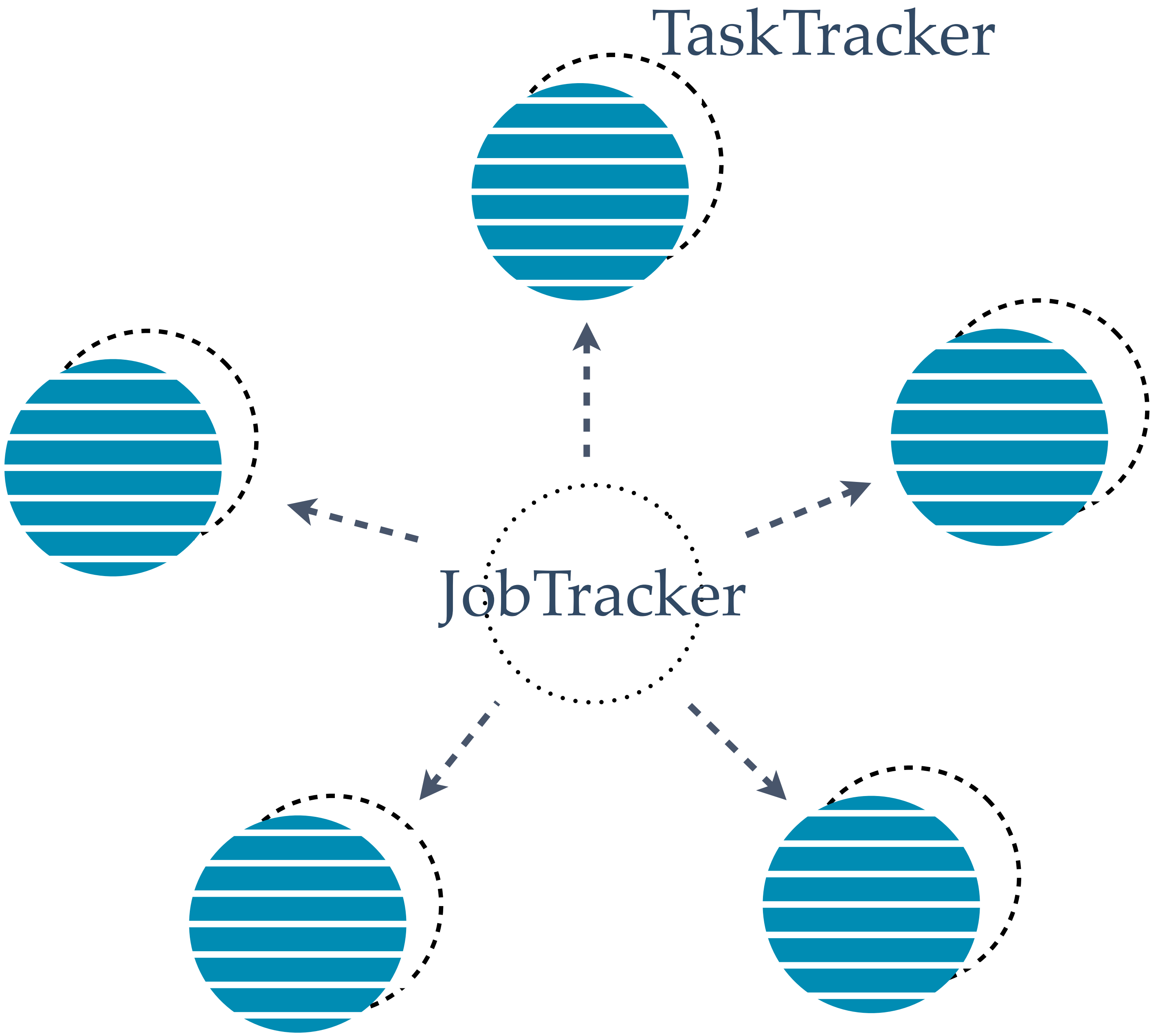
API cake

- libpq
- JDBC
- JPA

- Thrift
- Pelops, Hector
- Kundera, ?

Analytics in Cassandra

- @afex: “Cassandra + Pig (Hadoop) is very exciting. A 7 line script to analyze data from my entire cluster transparently, with no ETL? Yes, please”


TaskTracker
JobTracker

0.7 in November

- More control over replica placement
- Hadoop refinements
- Secondary indexes
- Online schema changes
- Large row support (> 2GB)
- Dynamic routing around slow nodes

When do you need Cassandra?

- Ian Eure: “If you’re deploying memcache on top of your database, you’re inventing your own ad-hoc, difficult to maintain NoSQL data store”

Not Only SQL

- Curt Monash: “**ACID-compliant transaction integrity** commonly costs more in terms of DBMS licenses and many other components of TCO (Total Cost of Ownership) than [scalable NoSQL]. Worse, it **can actually hurt application uptime**, by forcing your system to pull in its horns and stop functioning in the face of failures that a non-transactional system might smoothly work around. Other flavors of “complexity can be a bad thing” apply as well. Thus, **transaction integrity can be more trouble than it’s worth**.” [Curt’s emphasis]

More

- http://riptano.com/docs
- http://wiki.apache.org/cassandra/ArticlesAndPresentations
- http://wiki.apache.org/cassandra/ArchitectureInternals