Data Serving Systems in Cloud Computing Platforms

Sudipto Das

eXtreme Computing Group (XCG),

Microsoft Research (MSR)

Day 1 Afternoon Session

RETHINKING EVENTUAL CONSISTENCY

Definition: Eventual Consistency

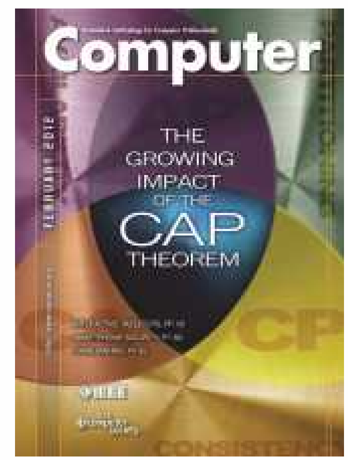
In a replicated database, updates arrive in different orders at different copies of a data item,

but eventually

the copies converge to the same value.

Eventual Consistency is All the Rage

- Origin: Thomas' majority consensus algorithm, published in 1979 (ACM TODS).
- Was used in Grapevine (PARC, early 1980's) and in numerous systems since then.
- Doug Terry et al. coined the term in a 1994 Bayou paper
- Werner Vogels at Amazon promoted it in Dynamo (2007)
- Cover topic of February 2012 IEEE Computer



Despite today's hype

- Most of what we'll say was known in 1995
- There are many published surveys
  - But this talk has a rather different spin
- We'll often cite old references to remind you where the ideas came from

Correctness Goal

- Ideally, replication is transparent

In the world of transactions:

- One-Copy Serializability - The system behaves like a serial processor of transactions on a one-copy database
  [Attar, Bernstein, & Goodman, IEEE TSE 10(6), 1984]

In the world of operations:

- Linearizability - A system behaves like a serial processor of operations on a one-copy database
  [Herlihy & Wing, ACM TOPLAS 12(3), 1990]

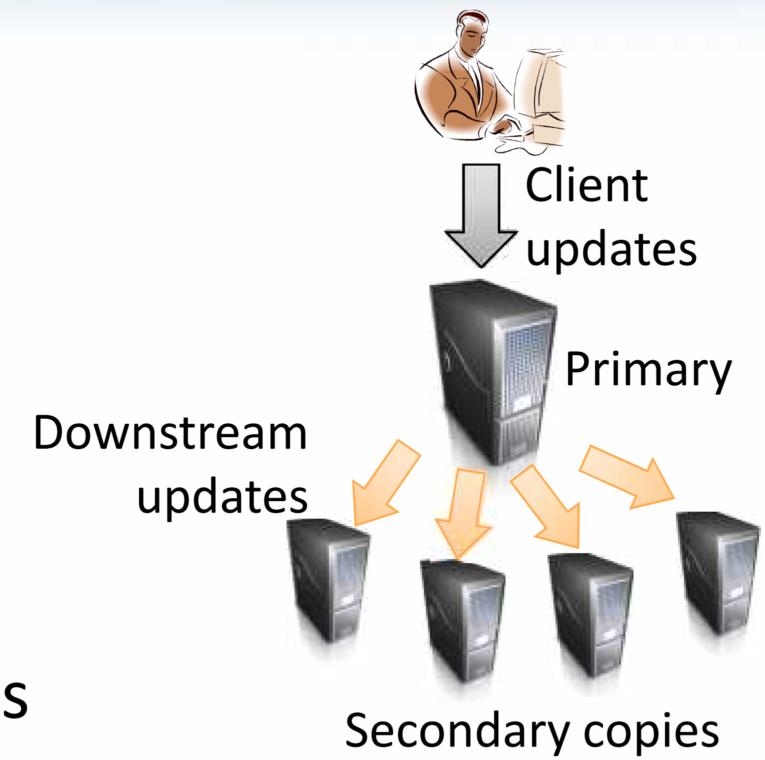
Nice Goal If You Can Get It

- But you can't in many practical situations

- Let's review the three main types of solutions
  - Primary Copy
  - Multi-Master
  - Consensus Algorithms

Primary Copy

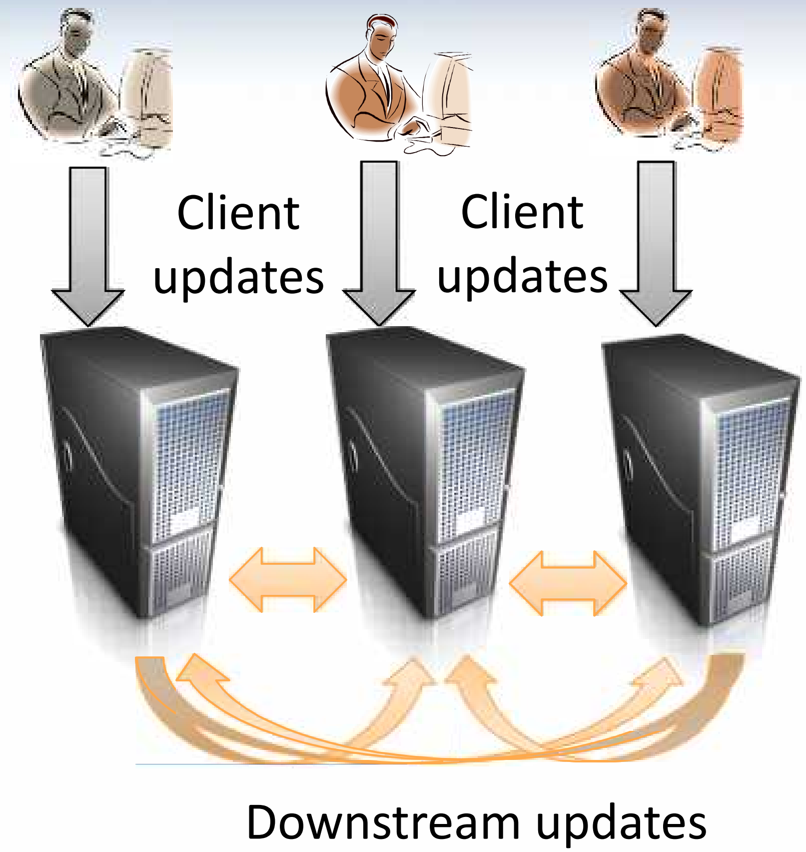
- Only the primary copy is updatable by clients
- Updates to the primary flow downstream to secondaries
- What if there's a network partition?
- Clients that can only access secondaries can't run updates



[Alsberg & Day, ICSE 1976] [Stonebraker & Neuhold, Berkeley Workshop 1979]

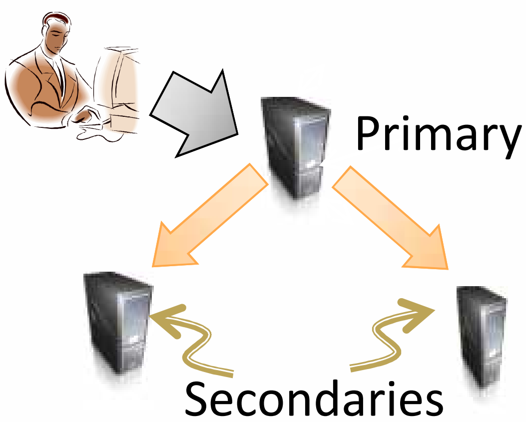
Multi-Master

- Copies are independently updatable
- Conflicting updates on different copies are allowed
- Doesn't naturally support 1SR.
- To ensure eventual consistency or linearizability of copies:
  - Either updates are designed to be commutative
  - Or conflicting updates are detected and merged
- "The partitioned DB problem" in late 1970's.
- Popularized by Lotus Notes, 1989

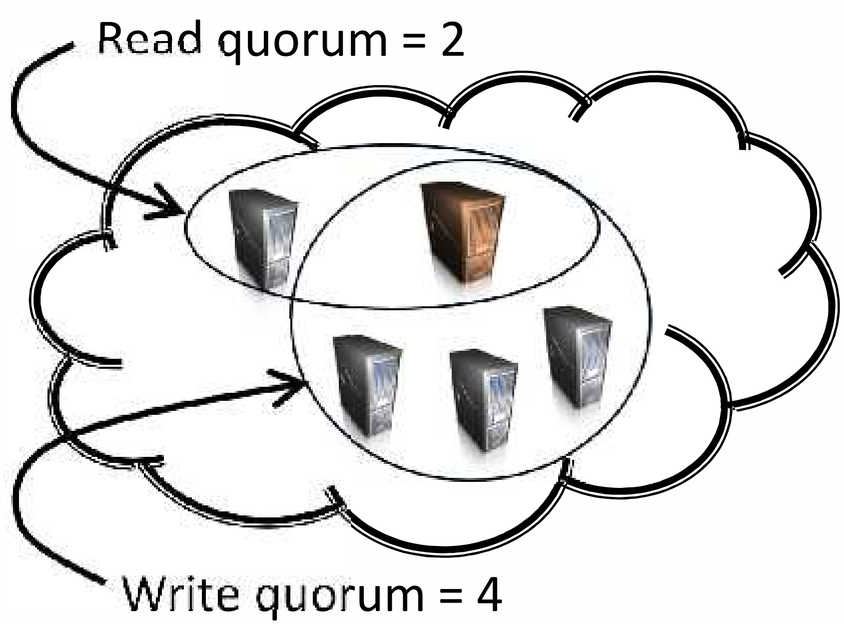


Consensus Algorithms

- Copies can be a replicated-state machine
  - Essentially, a serial processor of operations
  - Can be primary-copy or multi-master
- Uses quorum consensus to achieve 1SR or linearizability.
  - Ensures conflicting ops access at least one copy in common



Each downstream update is applied to a quorum of secondaries



The CAP Theorem

- You can have only two of Consistency-of-Replicas, Availability, and Partition-Tolerance
  - Can get C & A, if there's no partition
  - Can get C & P but only one partition can accept updates
  - Can get A & P, but copies in different partitions won't be consistent

Conjecture by [Brewer, PODC 2000]
Proved by [Gilbert & Lynch, SIGACT News 33(3) 2002]

This Isn't Exactly News

- "Partitioning - When communication failures break all connections between two or more active segments of the network ... each isolated segment will continue ... processing updates, but there is no way for the separate pieces to coordinate their activities. Hence ... the database ... will become inconsistent. This divergence is unavoidable if the segments are permitted to continue general updating operations and in many situations it is essential that these updates proceed."
- [Rothnie & Goodman, VLDB 1977]
- So the CAP theorem isn't new, but it does focus attention on the necessary tradeoff

Can we do better than Eventual Consistency?

- There have been many attempts at defining stronger but feasible consistency
  - Parallel snapshot isolation
  - Consistent prefix
  - Monotonic reads
  - Timeline consistency
  - Linearizability
  - Eventually consistent transactions
  - Causal consistency
  - Causal+ consistency
  - Bounded staleness
  - Monotonic writes
  - Read-your-writes
  - Strong consistency

It's Confusing

- We'll try to eliminate the confusion by
  - Characterizing consistency criteria
  - Describing mechanisms that support each one
  - And summarizing their strengths and weaknesses

Disclaimer

- There are many excellent surveys of replication
  - We don't claim ours is better, just different

- S.B. Davidson, H. Garcia-Molina, D. Skeen: Consistency in Partitioned Networks. ACM Computing Surveys. Sept. 1985
- S-H Son: Replicated data management in distributed database systems, SIGMOD Record 17(4), 1988.
- Y. Saito, M. Shapiro: Optimistic replication. ACM Comp. Surveys. Jan. 2005
- P. Padmanabhan, et al.: A survey of data replication techniques for mobile ad hoc network databases. VLDB Journal 17(5), 2008
- D.B. Terry: Replicated Data Management for Mobile Computing. Morgan & Claypool Publishers 2008
- B. Kemme, G. Alonso: Database Replication: a Tale of Research across Communities. PVLDB 3(1), 2010
- B. Kemme, R. Jiménez-Peris, M. Patiño-Martínez: Database Replication. Morgan & Claypool Publishers 2010

More Disclaimers

- There's a huge literature on replication.
  - Please tell us if we missed something important
- We'll cover replication mechanisms in database systems, distributed systems, programming languages, and computer-supported cooperative work
  - We won't cover mechanisms in computer architecture

Preview of the Design Space

- Multi-master is designed to handle partitions
- With primary copy, during a partition
  - Majority quorum(x) = partition with a quorum of x's copies
  - Majority quorum can run updates and satisfy all correctness criteria
  - Minority quorum can run reads but not updates, unless you give up on consistency
- So an updatable minority quorum is just like multi-master

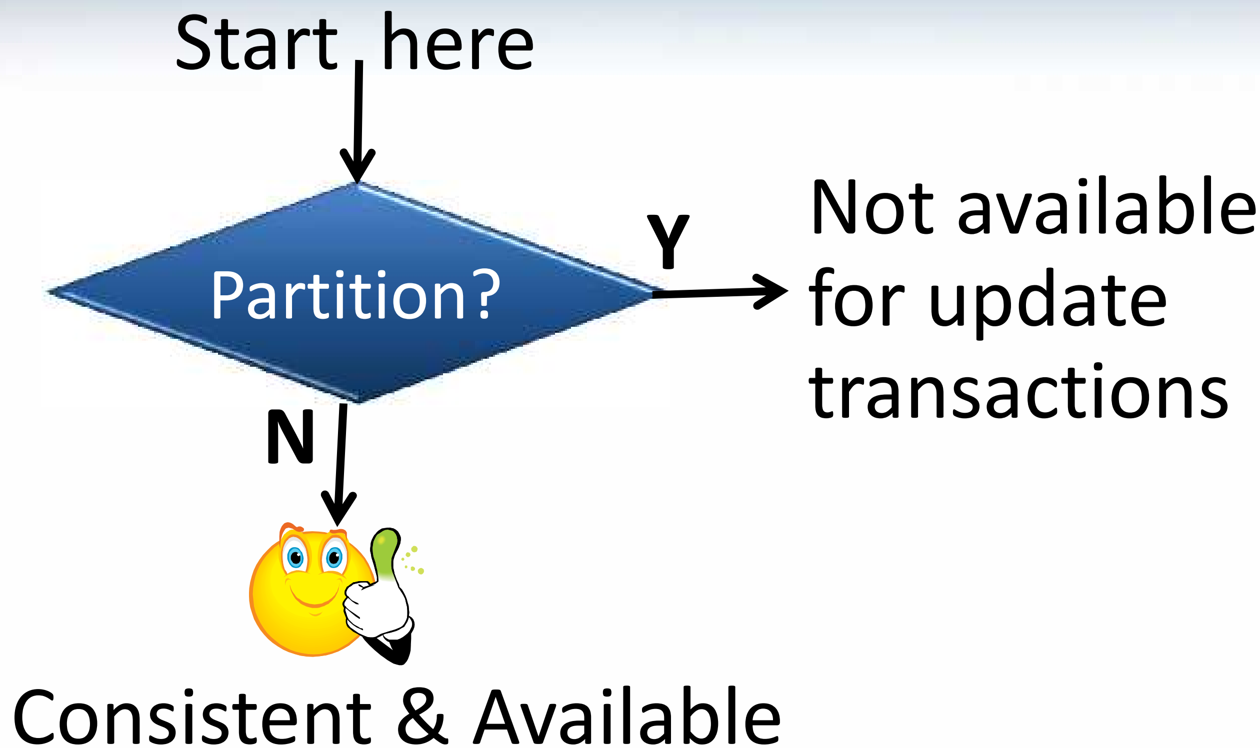
Preview of the Design Space (2)

- Eventual consistency – there are many good ways to achieve it
- For isolation and session goals, the solution space is more complex
  - Strengthens consistency, but complicates programming model
  - Improves availability, but not clear by how much
  - If a client rebinds to another server, ensuring these goals entails more expense, if they're attainable at all.
  - No clear winner

Bottom Line

- App needs to cope with arbitrary states during a partition
- Offer a range of isolation and session guarantees and let the app developer choose among them
  - Possibly worthwhile for distributed systems experts
  - Need something simpler for "ordinary programmers"
- Encapsulate solutions that offer good isolation for common scenarios
  - Use data types with commutative operations
  - Convergent merges of non-commutative operations
  - Scenario-specific classes

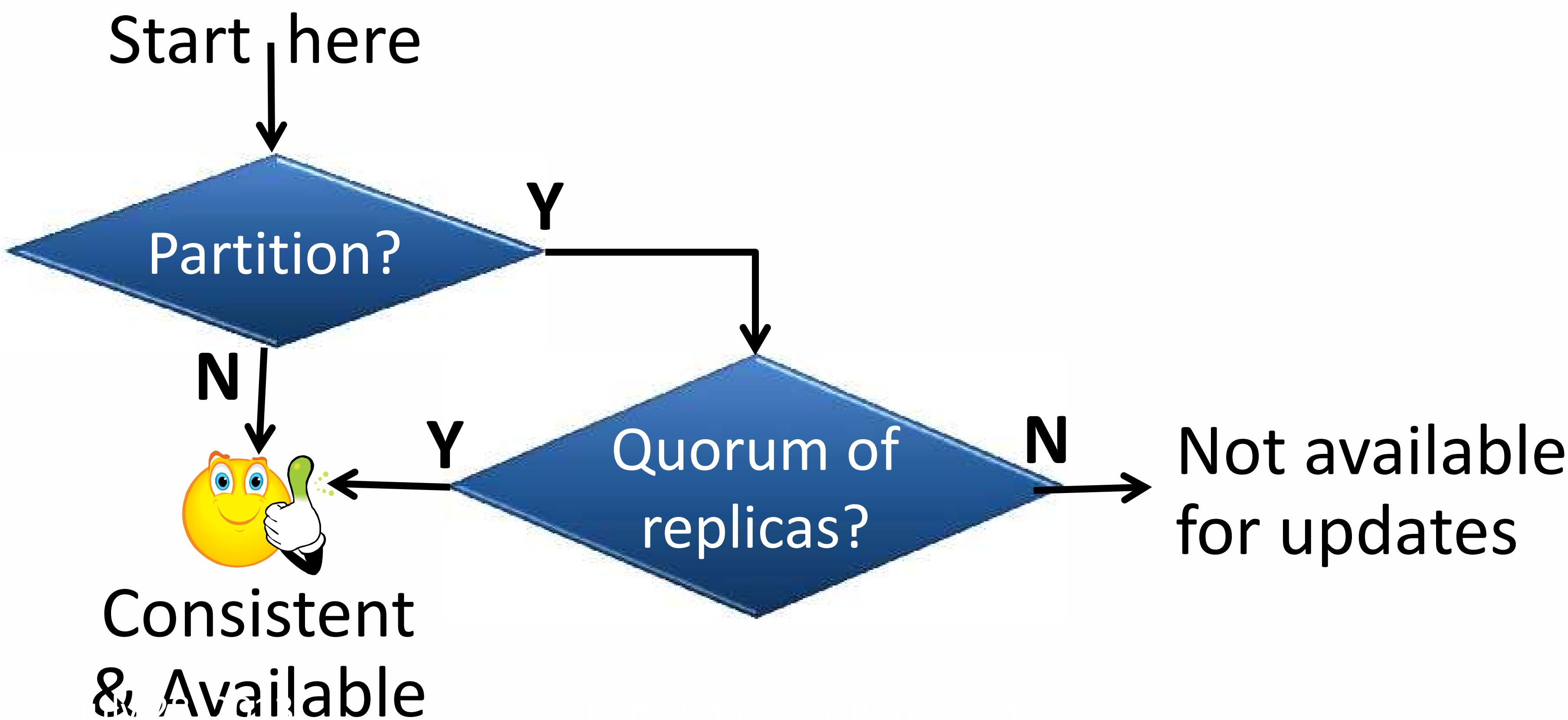
Organize Taxonomy by a Flowchart



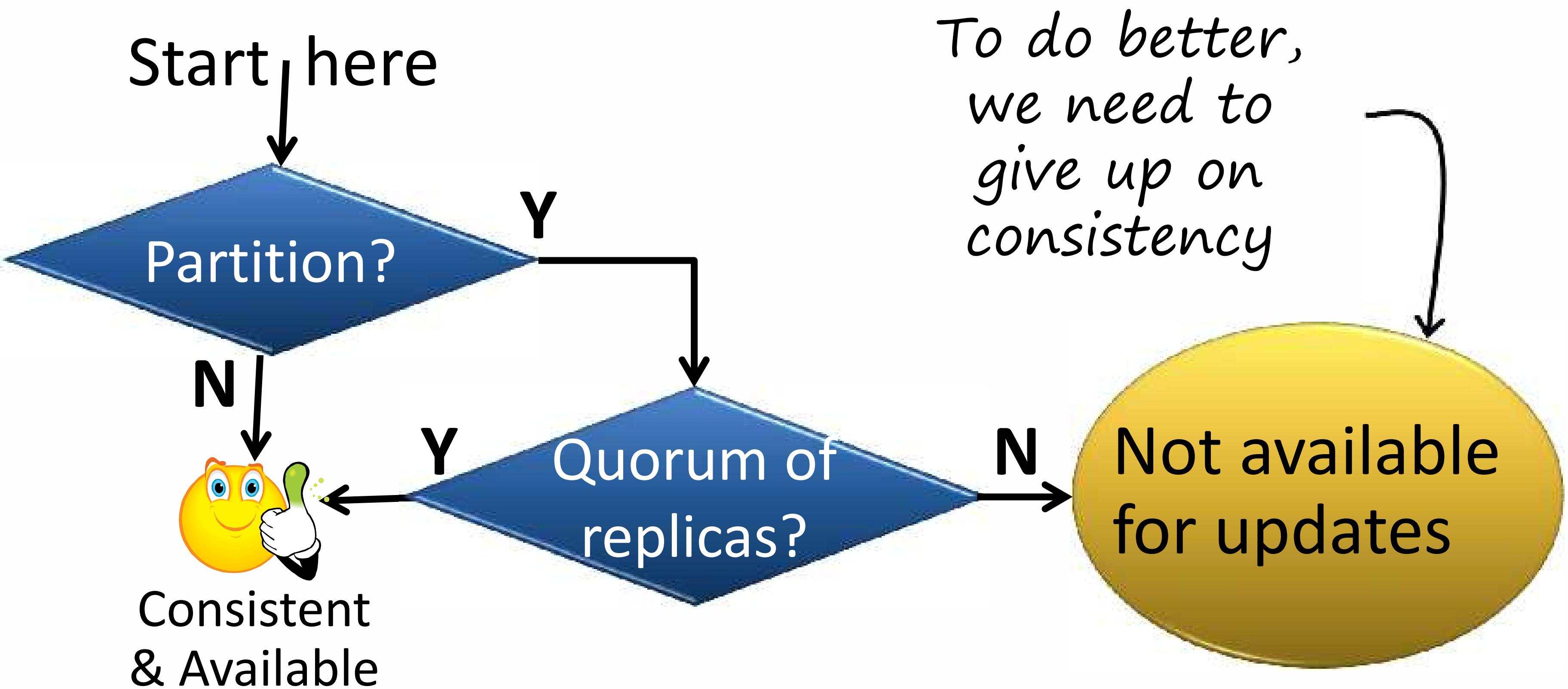
- We'll start with the world of operations, and then look at the world of transactions

Can Have One Writable Partition

- The partition with a quorum of replicas can run writes



What to do about the bad case?



Eventual Consistency

- Eventual consistency is one popular proposal
  - The copies will be identical … someday
  - App still needs to handle arbitrary intermediate states
- How to get it
  - Commutative downstream operations
  - Mergeable operations
  - Vector clocks

Commutative Downstream Updates

Thomas' Write Rule:

- [Thomas, ACM TODS 4(2), 1979]
- Assign a timestamp to each client write operation
- Each copy of *x* stores timestamp(last-write-applied)
- Apply downstream-write(*x*) only if downstream-write(*x*).timestamp > *x*.timestamp
- So highest-timestamp wins at every copy

Downstream writes arrive in this order

W(X=40), TS:1
W(X=70), TS:5
W(X=30), TS:3

Final value:
X=70, TS:5

Commutative Downstream Operations (2)

Pros

- Updates can be applied anywhere, anytime
- Downstream updates can be applied in any order after a partition is repaired

Cons

- Doesn't solve the problem of ordering reads & updates
- For fairness, requires loosely-synchronized clocks

Commutative Downstream Updates (3)

Convergent & Commutative Replicated Data Types

- [Shapiro et al., INRIA Tech. Report, Jan 2011]
- Set operations add/remove don't commute,
- [add(E), add(E), remove(E)] $\neq$ [add(E), remove(E), add(E)]
- But for a counting set, they do commute
  - Each element E in set S has an associated count
  - Add(set S, element E) increments the count for E in S.
  - Remove(S, E) decrements the count

Commutative Downstream Operations (4)

Pros

- Updates can be applied anywhere, anytime
- Downstream updates can be applied in any order after a partition is repaired

Cons

- Constrained, unfamiliar programming model
  - Doesn’t solve the problem of ordering reads & updates
- Some app functions need non-commutative updates

Custom Merge Operations

- Custom merge procedures for downstream operations whose client operations were not totally ordered.
  - Takes two versions of an object and creates a new one
- For eventual consistency, merge must be commutative and associative
- Notation: $M(O_2, O_1)$ merges the effect of O_2 into O_1
- Commutative: $O_1 \cdot M(O_2, O_1) \equiv O_2 \cdot M(O_1, O_2)$
- Associative: $M(O_3, O_1 \cdot M(O_2, O_1)) \equiv M(M(O_3, O_2) \cdot O_1)$
- [Ellis & Gibbs, SIGMOD 1989]

Custom Merge Operations (cont'd)

Pros

- Enables concurrent execution of conflicting operations without the synchronization expense of total-ordering

Cons

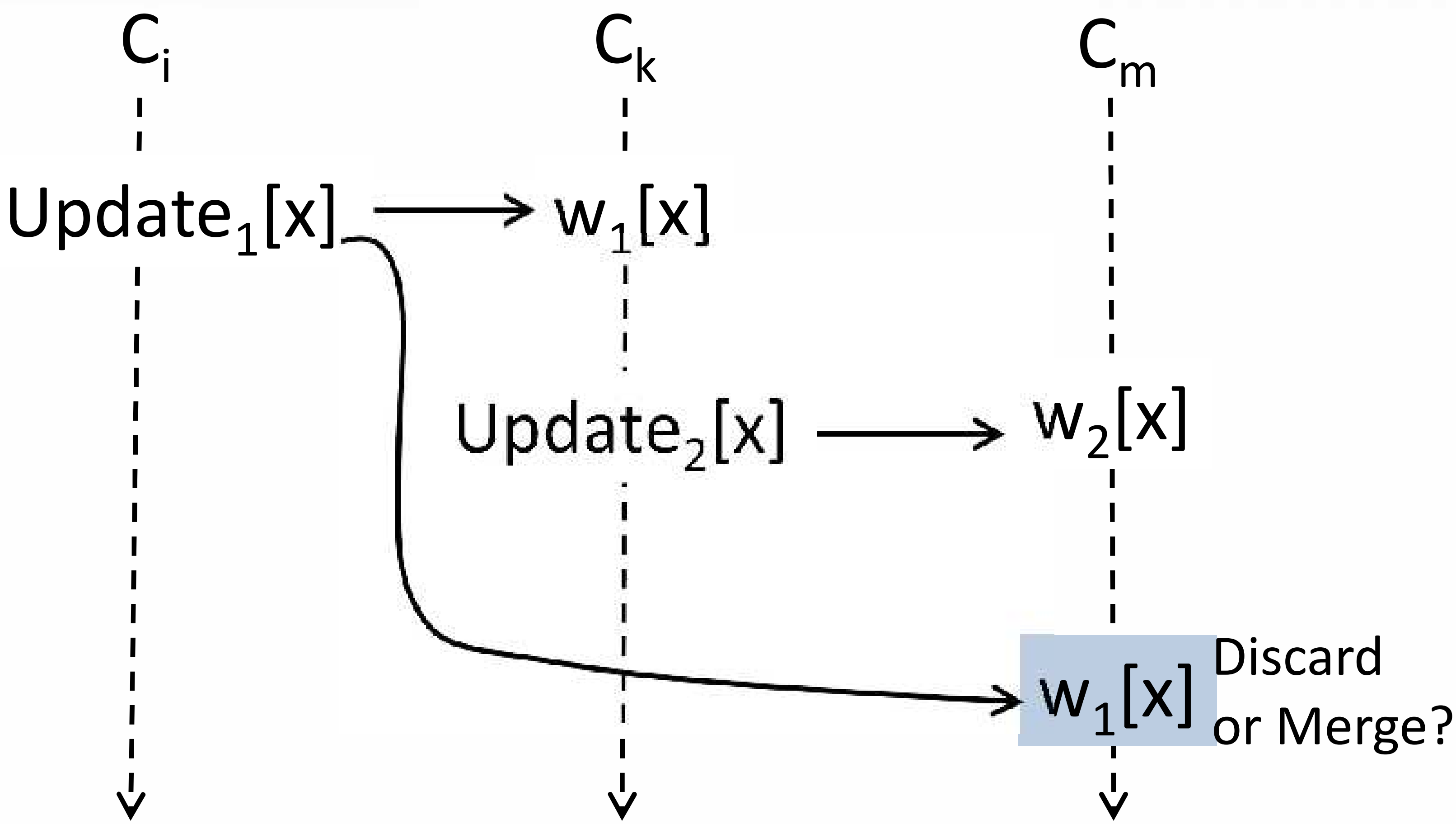
- Requires application-specific logic that's hard to generalize

Vector Clocks tell us the merging order

- In multi-master, each copy assigns a monotonically increasing version number to each client update
- Vector clock is an array of version numbers, one per copy
  - Identifies the set of updates received or applied
- Use it to identify the state that a client update depends on and hence overwrote
  - If two updates conflict but don't depend on one another, then merge them.

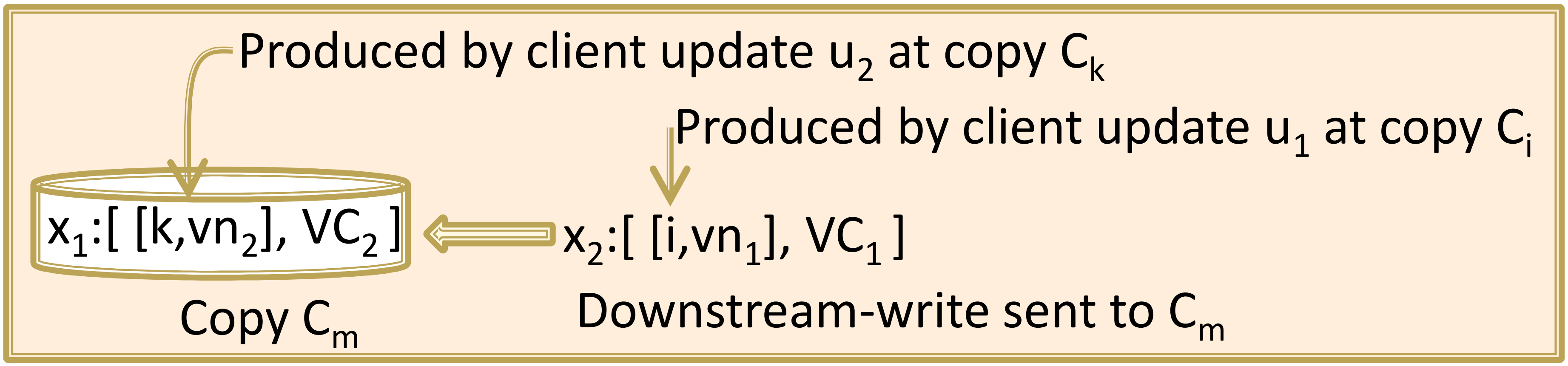
- [Fischer & Michael, PODS 1982]
- [Parker et al., IEE TSE 1983]
- [Wuu & Bernstein, PODC 1984]

Problem: Discard or Merge?



Vector Clocks(2)

- A vector clock can be used to identify the state that a client update depends on (“made-with knowledge”)



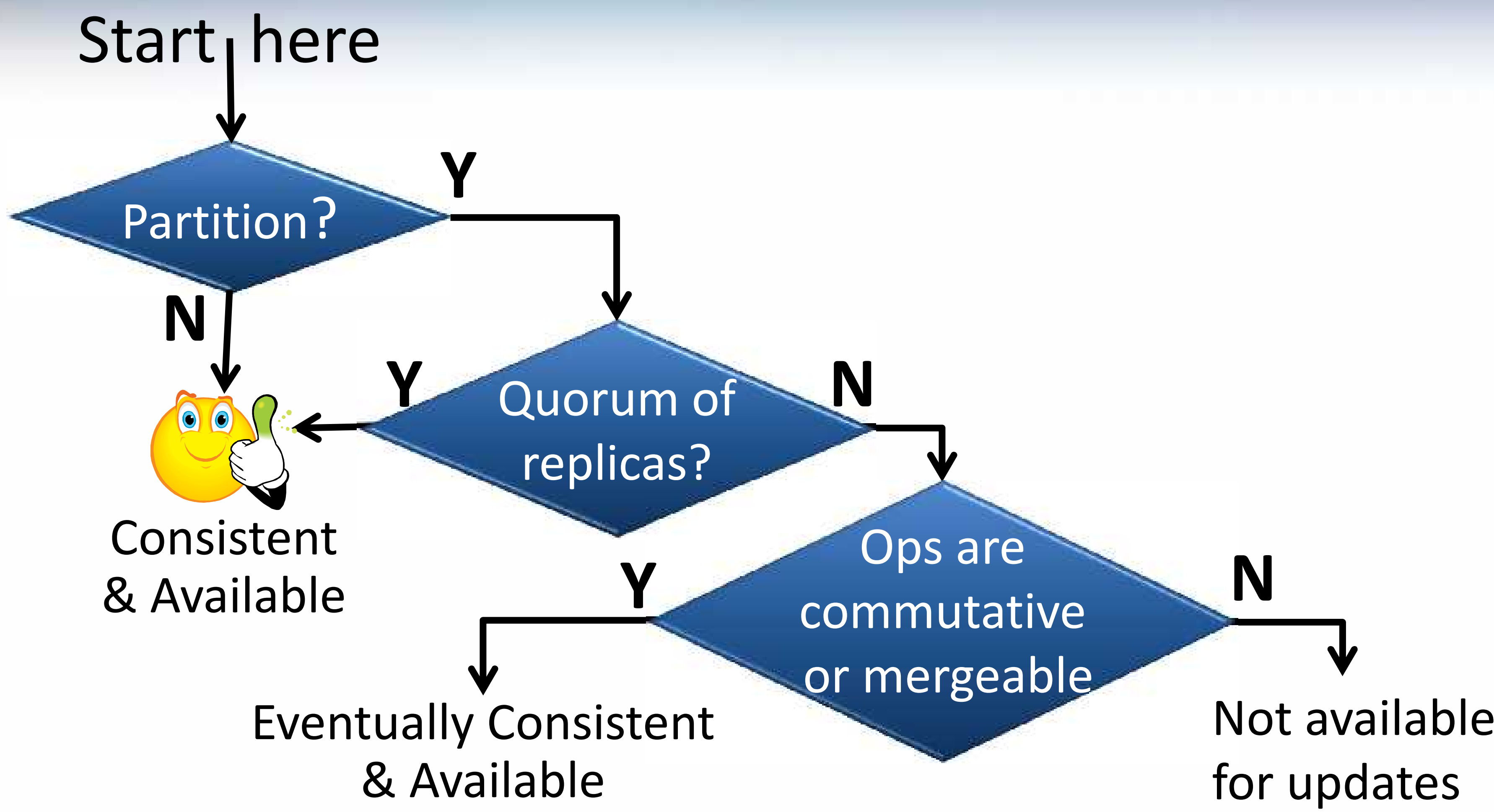
  - If $VC_1[k] \geq vn_2$, then x_2 was “made from” x_1 & should overwrite it
  - If $VC_2[i] \geq vn_1$, then x_1 was “made from” x_2, so discard x_2
  - Else the updates should be reconciled

[Ladin et al., TOCS, 1992]
[Malkhi & Terry, Dist. Comp. 20(3), 2007]

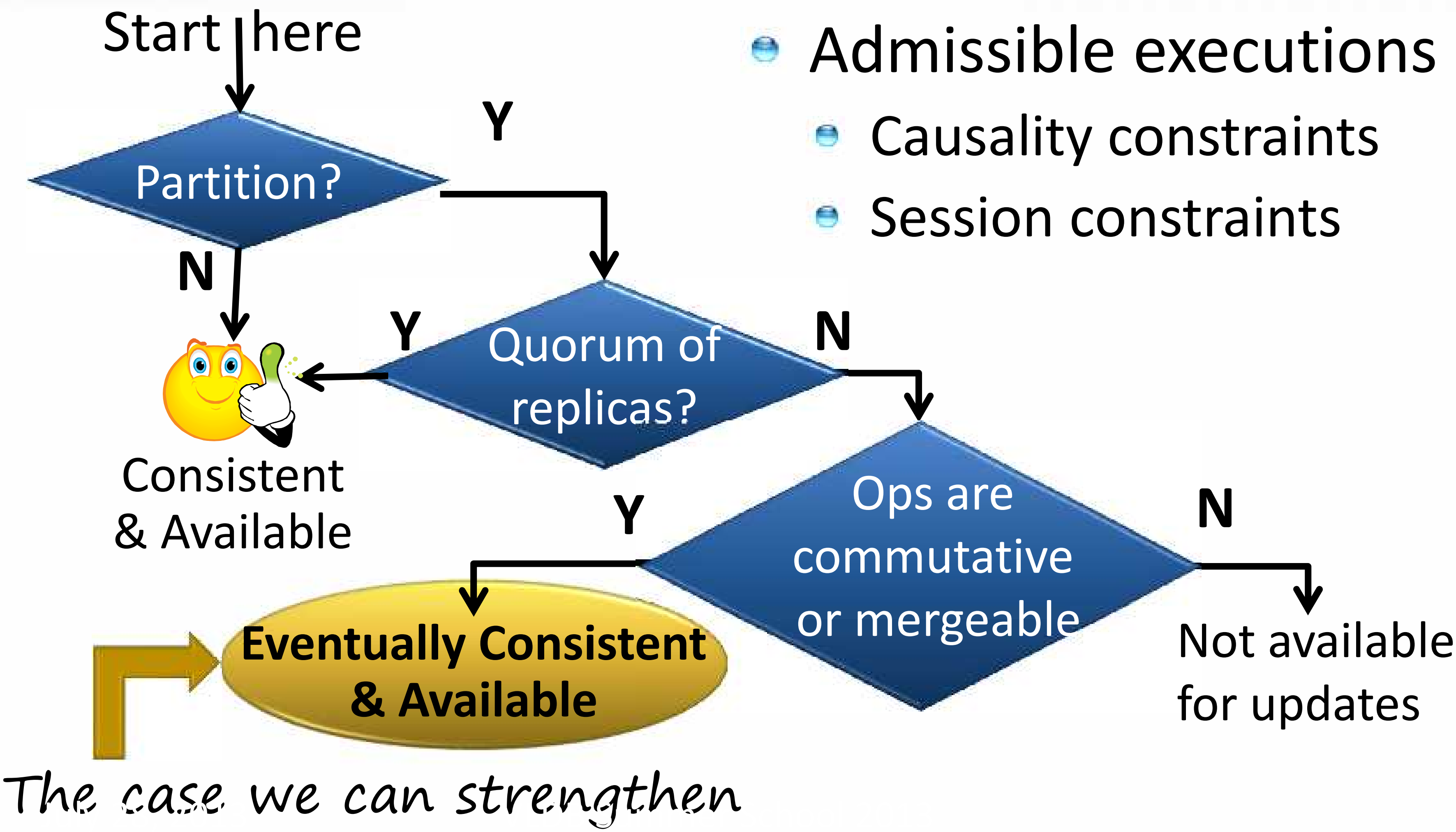
Another Use of Vector Clocks

- A copy can use it to identify the updates it has received
  - When it syncs with another copy, they exchange vector clocks to tell each other which updates they already have.
- Avoids shipping updates the recipient has already seen
- Enables a copy to discard updates that it knows <u>all</u> other copies have seen

In the Operation World



Strengthening Eventual Consistency

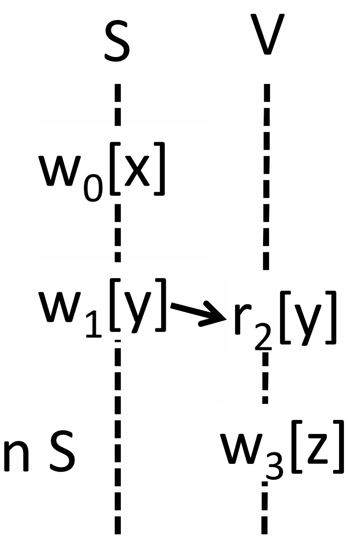


- Admissible executions
  - Causality constraints
  - Session constraints

Causal Consistency

Definition – The sequence of operations on each replica is consistent with session order and reads-from order.

- Example: User 1 stores a photo P and a link L to it.
  If user 2 reads the link, then she'll see the photo.
- Causality imposes write-write orders



Causal relationships:

- **WW Session order:** $w_1[y]$ executes after $w_0[x]$ in session S
- **WR Session order**: $w_3[z]$ executes after $r_2[y]$ in session V
- **Reads-from order**: $r_2[y]$ in session V reads from $w_1[y]$ in session S
- **Causality is transitive:** Hence, $w_0[x]$ causally precedes $w_3[z]$

[Lamport, CACM 21(7), 1978]

Causal Consistency (2)

- If all atomic operations preserve database integrity, then causal consistency with eventual consistency may be good enough
  - Store an object, then a pointer to the object
  - Assemble an order and then place it
  - Record a payment (or any atomically-updatable state)
- Scenarios where causal consistency isn't enough
  - Exchanging items: Purchasing or bartering require each party to be credited and debited atomically
  - Maintaining referential integrity: One session deletes an object O while another inserts a reference to O

Implementing Causal Consistency

- Enforce it using dependency tracking and vector clocks
- COPS: Causality with convergent merge [Lloyd et al., SOSP 2011]
  - Assumes multi-master replication
  - Session context (dependency info) = <data item, version#> of the last items read or of the last item written.
  - Each downstream write includes its dependent operations.
  - A write is applied to a copy after its dependencies are satisfied
  - Merge uses version vectors
  - With additional dependency info, it can support snapshot reads
  - Limitation: No causal consistency if a client rebinds to another replica due to a partition

Session Constraints

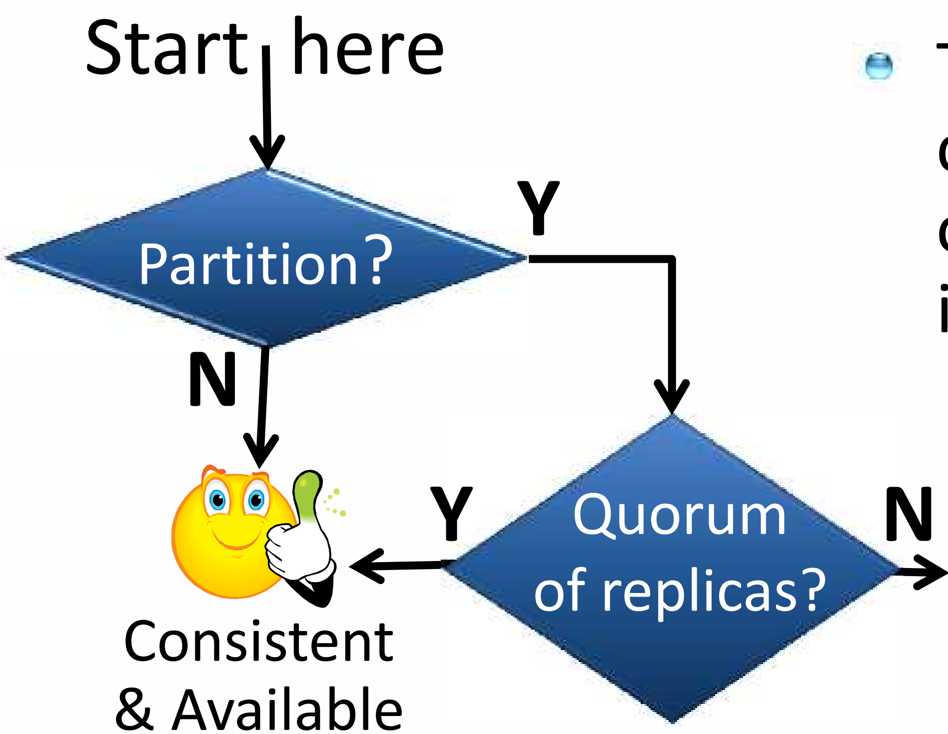
- **Read your writes** – a read sees all previous writes
- **Monotonic reads** – reads see progressively later states
- **Monotonic writes** – writes from a session are applied in the same order on all copies
- **Consistent prefix** – a copy's state only reflects writes that represent a prefix of the entire write history
- **Bounded staleness** – a read gets a version that was current at time *t* or later

[Terry et al., PDIS 1994]

Mechanisms for Session Constraints

- Client session maintains IDs of reads and writes
  - ✓ Accurate representation of the constraints
  - ☹ High overhead per-operation
- Client session maintains vector clocks for the last item read or written
  - ✓ Compact representation of the constraints
  - ☹ Conservative

In the Transaction World



- The operation world ignores transaction isolation
- To get the benefits of commutative or mergeable operations, need a weaker isolation level

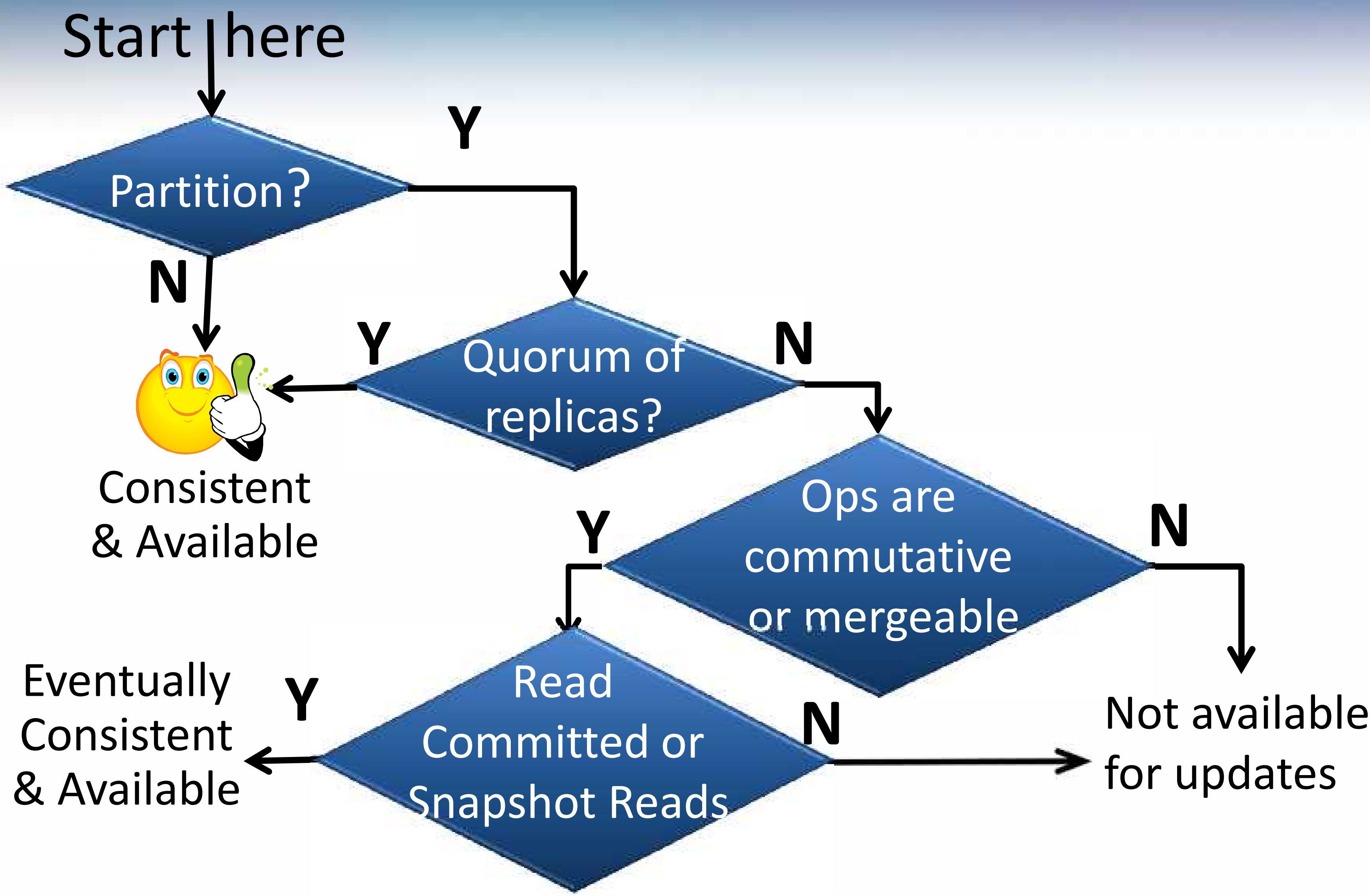
Common weaker isolation levels

- Read committed
  - Transaction reads committed values
- Snapshot reads
  - Transaction reads committed values that were produced by a set of committed transactions
  - All of a transaction's updates must be installed atomically to ensure the writeset is consistent in the minority partition

Is Weaker Isolation Acceptable?

- People do it all the time for better performance
  - Throughput of Read-Committed is 2.5x to 3x that of Serializable
- Weaker isolation produces errors. Why is this OK?
- No one knows, but here are some guesses:
  - DB's are inconsistent for many other reasons.
    - Bad data entry, bugs, duplicate txn requests, disk errors, ….
  - Maybe errors due to weaker isolation levels are infrequent
  - When DB consistency matters a lot, there are external controls.
    - People look closely at their paychecks
    - Financial information is audited
    - Retailers take inventory periodically

In the Transaction World



Other Admissibility Constraints

- Admissible executions
  - Causality constraints
  - Session constraints
  - Isolation constraints
    - **RedBlue Consistency [Li et al., OSDI 2012]**
    - **1-SR, Read-committed, Snapshot Isolation**
    - **Parallel Snapshot Isolation [Sovran et al, SOSP 2011]**
    - **Concurrent Revisions [Burckhardt et al., ESOP 2012]**

RedBlue Consistency

- *Blue* operations commute with all other operations and can run in different orders on different copies.
- *Red* ones must run in the same order on all copies.
- Use a side-effect-free *generator* operation to transform a red operation to a blue one that is valid in all states
- Example
  - Deposit(acct, amt): acct.total = acct.total + amt
  - EarnInterest(acct): acct.total = acct.total * 1.02
  - Deposit is blue, EarnInterest is red
  - Transform EarnInterest into:
    - Interest = acct.total * 1.02 // runs locally at acct's copy
    - Deposit(acct, Interest)   // blue operation runs at all copies

[Li et al., OSDI 2012]

Snapshot Isolation (SI)

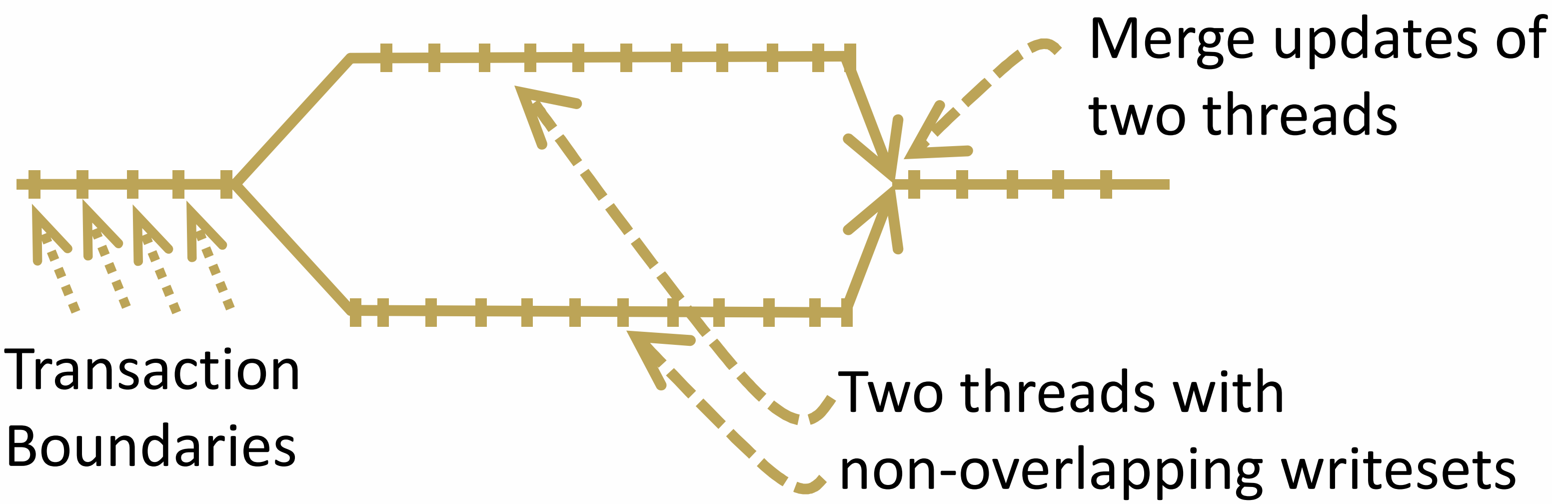
- The history is equivalent to one of this form:

$r_1[readset_1]$ $w_1[writeset_1]$	$r_4[readset_4]$ $w_4[writeset_4]$		
$r_2[readset_2]$ $w_2[writeset_2]$	$r_5[readset_5]$ $w_5[writeset_5]$	• • •	
$r_3[readset_3]$ $w_3[writeset_3]$	$r_6[readset_6]$ $w_6[writeset_6]$		
$ws_1 \cap ws_2 \cap ws_3 = \varnothing$	$ws_4 \cap ws_5 \cap ws_6 = \varnothing$		

- Benefit of SI: Don't need to test read-write conflicts

Parallel Snapshot Isolation (PSI)

- Parallel SI - Execution is equivalent to one that allows parallel threads with non-conflicting writesets running SI
- Allows a transaction to read stale copies



[Sovran, Power, Aguilera, & Li, SOSP 2011]

Example: Parallel SI

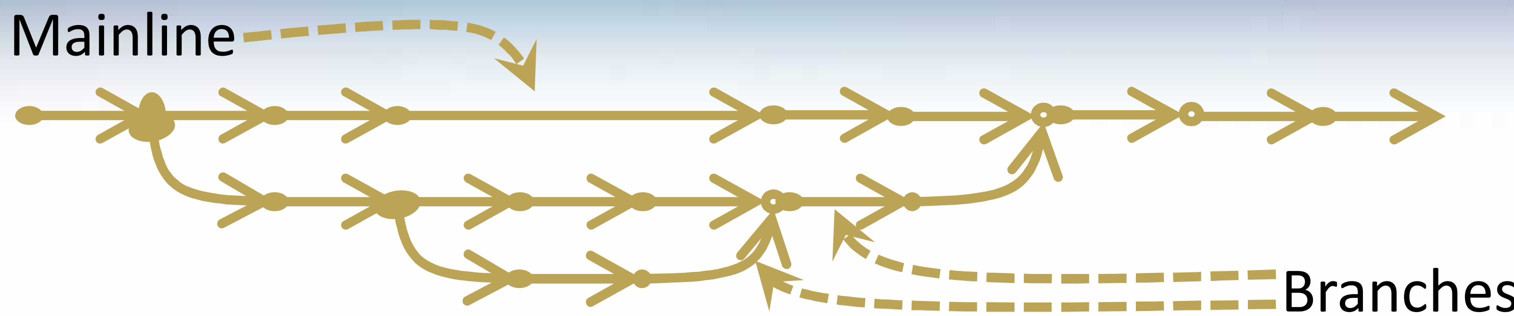
Site 1	Site 2
$r_1(x,y)$	$r_3(x,y)$
$w_1(x)$	$w_3(y)$
c_1	c_3
$r_2(x,y)$	$r_4(x,y)$
$w_2(x)$	$w_4(y)$
c_2	c_4
$w_3[y]$	$w_1[x]$
$w_4[y]$	$w_2[x]$

Site 1 has x's primary
Site 2 has y's primary

- A parallel SI execution may not be equivalent to a serial SI history
- Site 1 and Site 2 are each snapshot isolated.
- But the result is not equivalent to
  $T_1\ T_2\ T_3\ T_4$ or
  $T_3\ T_4\ T_1\ T_2$ or

T1	T2	• • •
T3	T4	

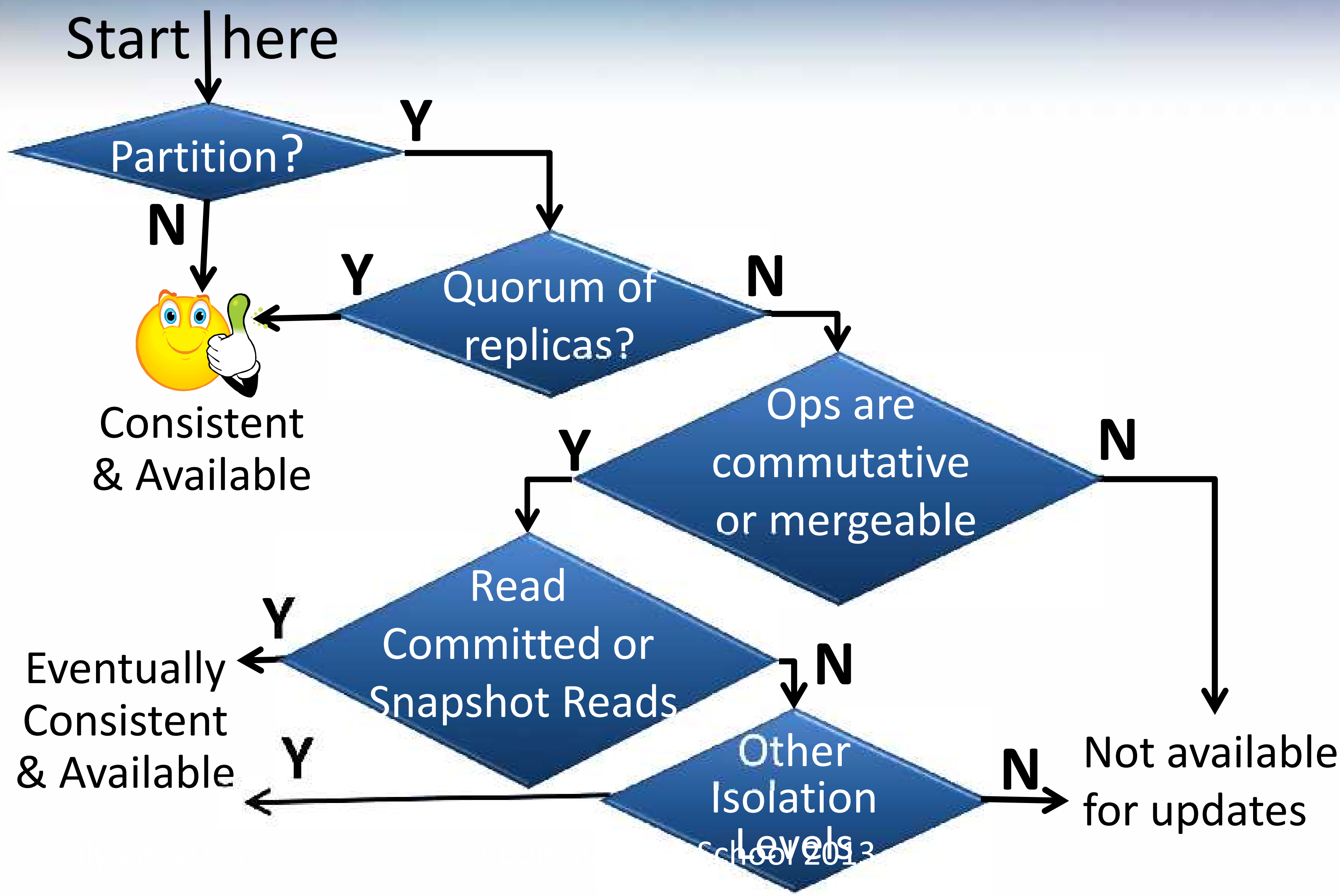
Concurrent Revisions



- Each arrow is an operation or transaction
- A fork defines a new private snapshot and a branch
- A join causes all updates on the branch to be applied
- Ops are pure reads or pure writes. Writes never fail.

[Burckhardt, et al., ESOP 2012]

In the Transaction World



RETURNING TO CAP …

Guidance for App Development

- If the system guarantees only eventual consistency, then be ready to read nearly arbitrary database states.
- Use commutative operations whenever possible.
  - System needn't totally order downstream writes, which reduces latency
- Else use convergent merges of non-commutative ops
  - Enables updates during partitioned operation and in multi-master systems

Guidance for Development (2)

- If availability and partition-tolerance are required, then consider strengthening eventual consistency with admissibility criteria
- If possible, use consistency-preserving operations, in which case causal consistency is enough
- Hard case for all admissibility criteria is rebinding a session to a different replica
  - Replica might be older or newer than the previous one it connected to.

Enforcing Admissibility in a Minority Partition

	Session maintains connection to server		Session migrates to another replica	
	Primary Copy or Quorum-based	**Multi-master**	**Primary Copy or Quorum-based**	**Multi-master**
Read-Your-Writes	✓	✓	✓ ?W	✓ ?W
Monotonic Writes	✓	✓	✓	✓ ?W
Bounded staleness	☹	☹	☹	☹
Consistent Prefix	✓	☹	✓	☹
Monotonic Reads	✓	✓	✓ ?R	✓ ?R
Causality	✓	✓	☹	☹

Writes disabled

?W: Only if the session caches its writes
?R: Only if the session caches its reads

Research Opportunity

- Encapsulate solutions that offer good isolation for common scenarios
  - Commutative Replicated Data Types
  - Convergent merges of non-commutative operations
  - Research: Scenario-specific design patterns
    - Overbooking with compensations
    - Queued transactions
    - • • •

Does this design space matter?

- Probably not to enterprise developers
- Spanner [OSDI 2012] “Many applications at Google … use Megastore because of its semi-relational data model and support for synchronous replication, despite its relatively poor write throughput.”
- Mike Stonebraker [blog@ACM, Sept 2010]:
  “No ACID Equals No Interest” for enterprise users

So Why Bother?

- The design space does matter to Einstein-level developers of high-value applications that need huge scale out.

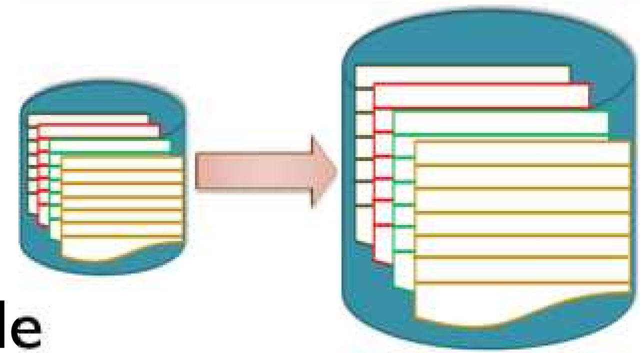
Summary

- Eventual consistency
  - Commutative operations
    - Thomas’ write rule
    - Convergent data types
  - Custom merge
    - Vector clocks
- Admissible executions
  - Causality constraints
  - Session constraints
    - Read your writes
    - Monotonic reads
    - Monotonic writes
    - Consistent prefix
    - Bounded staleness
  - Isolation constraints

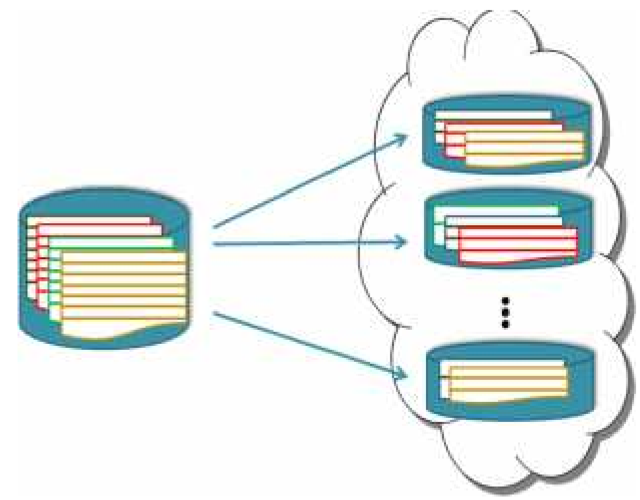
SCALE-OUT TRANSACTION PROCESSING

Two approaches to scalability

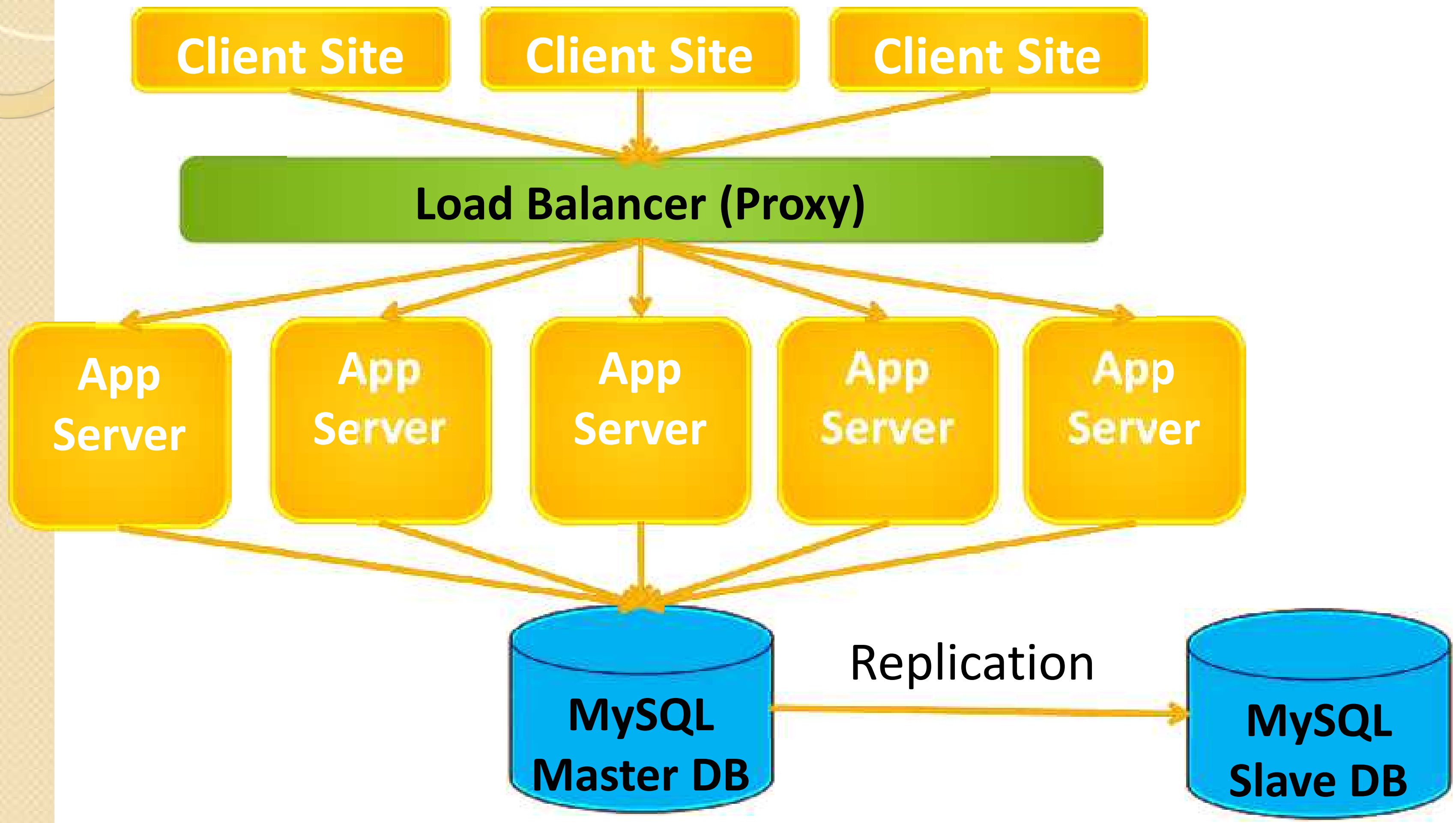
- **Scale-up**
  - Preferred in **classical enterprise** setting (RDBMS)
  - Flexible **ACID transactions**
  - Transactions access a single node



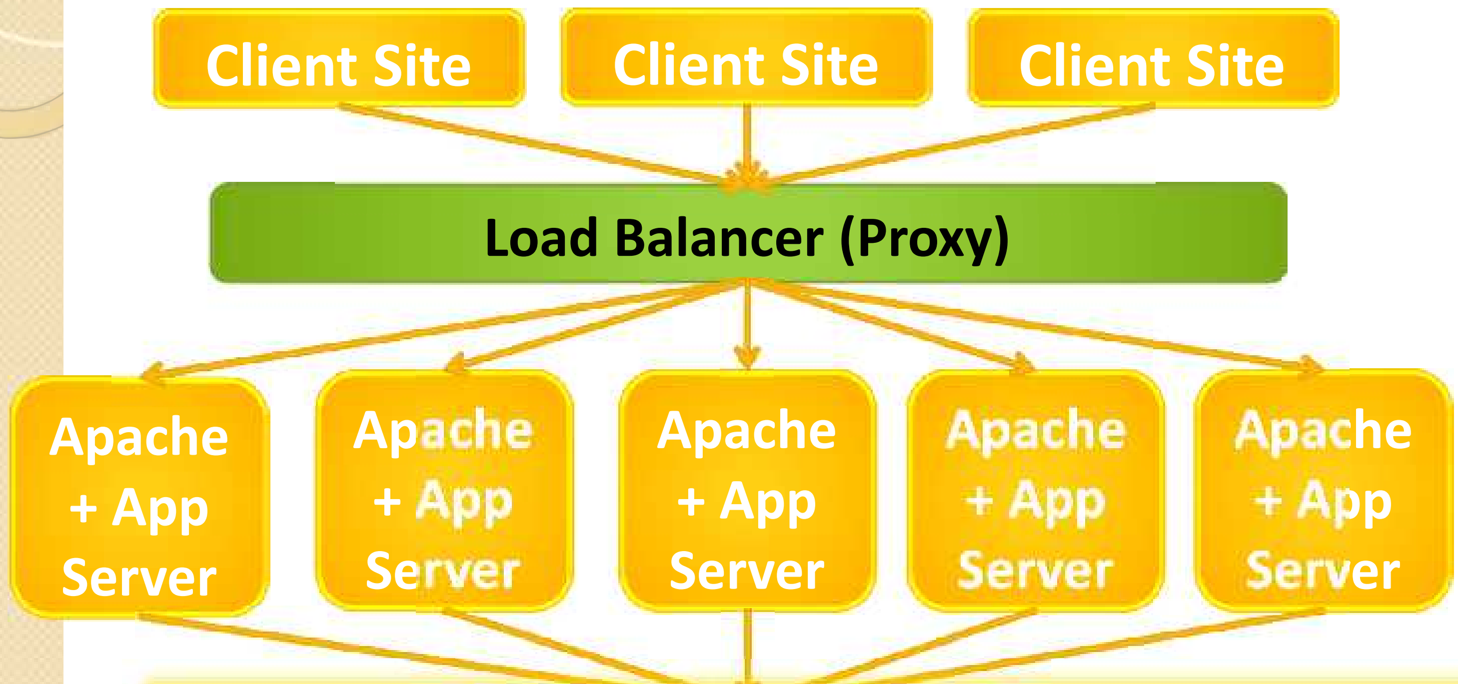
- **Scale-out**
  - **Cloud friendly** (Key value stores)
  - Execution at a single server
    - Limited functionality & guarantees
  - No **multi-row** or **multi-step** transactions



Scaling in the Cloud



Scaling in the Cloud



Scalable and Elastic, but limited consistency and operational flexibility

Blog Wisdom

- "If you want vast, on-demand scalability, you need a non-relational database." Since scalability requirements:
  - Can change very quickly and,
  - Can grow very rapidly.
- Difficult to manage with a single in-house RDBMS server.
- RDBMS scale well:
  - When limited to a single node, but
  - Overwhelming complexity to scale on multiple server nodes.

The "NoSQL" movement

- Initially used for: *"Open-Source relational database that did not expose SQL interface"*
- Popularly used for: "*non-relational, distributed data stores that often did not attempt to provide ACID guarantees*"
- Gained widespread popularity through a number of open source projects
  - HBase, Cassandra, Voldemort, MongDB, …
- Scale-out, elasticity, flexible data model, high availability

NoSQL has no relation with SQL

- Micheal Stonebraker [CACM Blog]

- Term heavily used (and abused)
- Scalability and performance bottleneck not inherent to SQL
  - Scalability, auto-partitioning, self-manageability can be achieved with SQL
- Different implementations of SQL engine for different application needs
- SQL provides flexibility, portability

No-SQL → Not Only SQL

- Recently renamed
- One size does not fit all
- Encompass a broad category of "structured" storage solutions
  - RDBMS is a subset
  - Key Value stores
  - Document stores
  - Graph database
- The debate on appropriate characterization continues

Why care about transactions?

```
confirm_friend_request(user1, user2)
{
begin_transaction();
        update_friend_list(user1, user2, status.confirmed);
        update_friend_list(user2, user1, status.confirmed);
end_transaction();
}
```

Simplicity in application design with ACID transactions

Sacrificing Consistency

Handle failures

```
confirm_friend_request_A(user1, user2) {
 try {
      update_friend_list(user1, user2, status.confirmed);
 } catch(exception e) {
      report_error(e);
      return;
 }
 try {
      update_friend_list(user2, user1, status.confirmed);
  } catch(exception e) {
      revert_friend_list(user1, user2);
      report_error(e);
      return;
  }
}
```

Sacrificing Consistency

Ensuring persistence

```
confirm_friend_request_B(user1, user2) {
  try{
     update_friend_list(user1, user2, status.confirmed);
  } catch(exception e) {
     report_error(e);
     add_to_retry_queue(operation.updatefriendlist, user1, user2, current_time());
  }

  try {
    update_friend_list(user2, user1, status.confirmed);
  } catch(exception e) {
   report_error(e);
   add_to_retry_queue(operation.updatefriendlist, user2, user1, current_time());
  }
}
```

```
confirm_friend_request_A(user1, user2) {
 try {
     update_friend_list(user1, user2, status.confirmed);
 } catch(exception e) {
     report_error(e);
     return;
 }
 try {
     update_friend_list(user2, user1, status.confirme
  } catch(exception e) {
     revert_friend_list(user1, user2);
     report_error(e);
     return;
  }
}

confirm_friend_request_B(user1
 try{
    update_friend_list(user1                    ied);
 } catch(exception e) {
    report_error(e);
    add_to_retry_que           riendlist, user1, user2, current_time());
 }

 try {
   update_fr                  1, status.confirmed);
 } catch(e
   report_
   add_to_re            eration.updatefriendlist, user2, user1, current_time());
 }
}
```

It gets too complicated with reduced consistency guarantees

Scale-out Transaction Processing

- Transactions on co-located data
- Transactions on distributed data

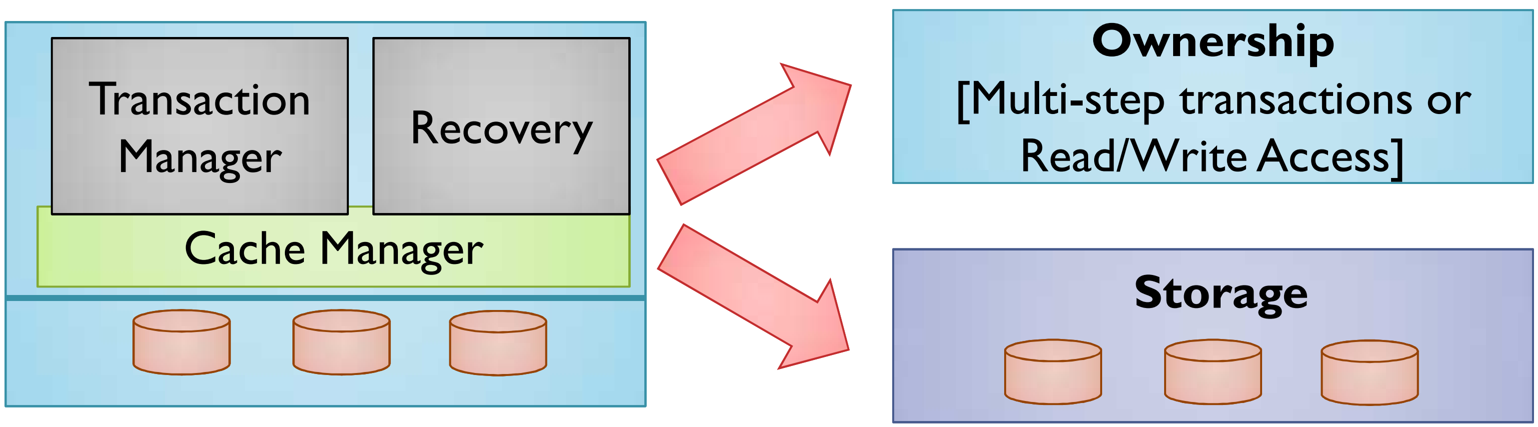
DESIGN PRINCIPLES (REVISITED)

Design Principles

- **Separate System** and **Application** State
  - **System metadata** is critical but small
  - **Application data** has varying needs
  - Separation allows use of different class of protocols

Design Principles

- **Decouple Ownership from Data Storage**
  - Ownership is exclusive read/write access to data
  - Decoupling allows lightweight ownership migration

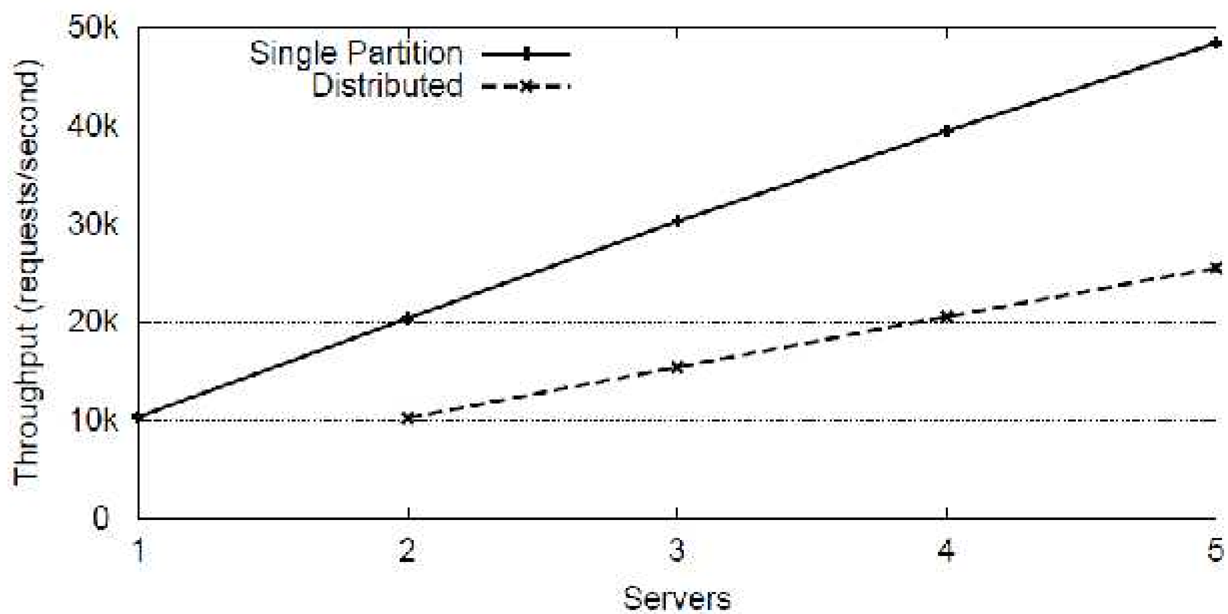


Classical DBMSs

Decoupled ownership and Storage

Design Principles

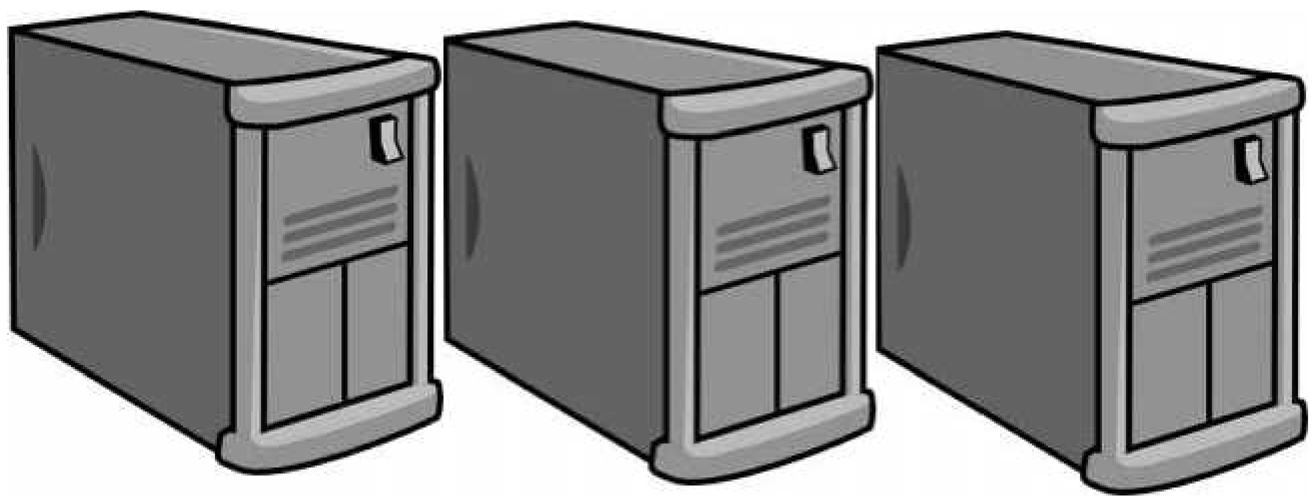
- **Limit most interactions to a single node**
  - Allows **horizontal scaling**
  - **Graceful degradation** during failures
  - No distributed synchronization



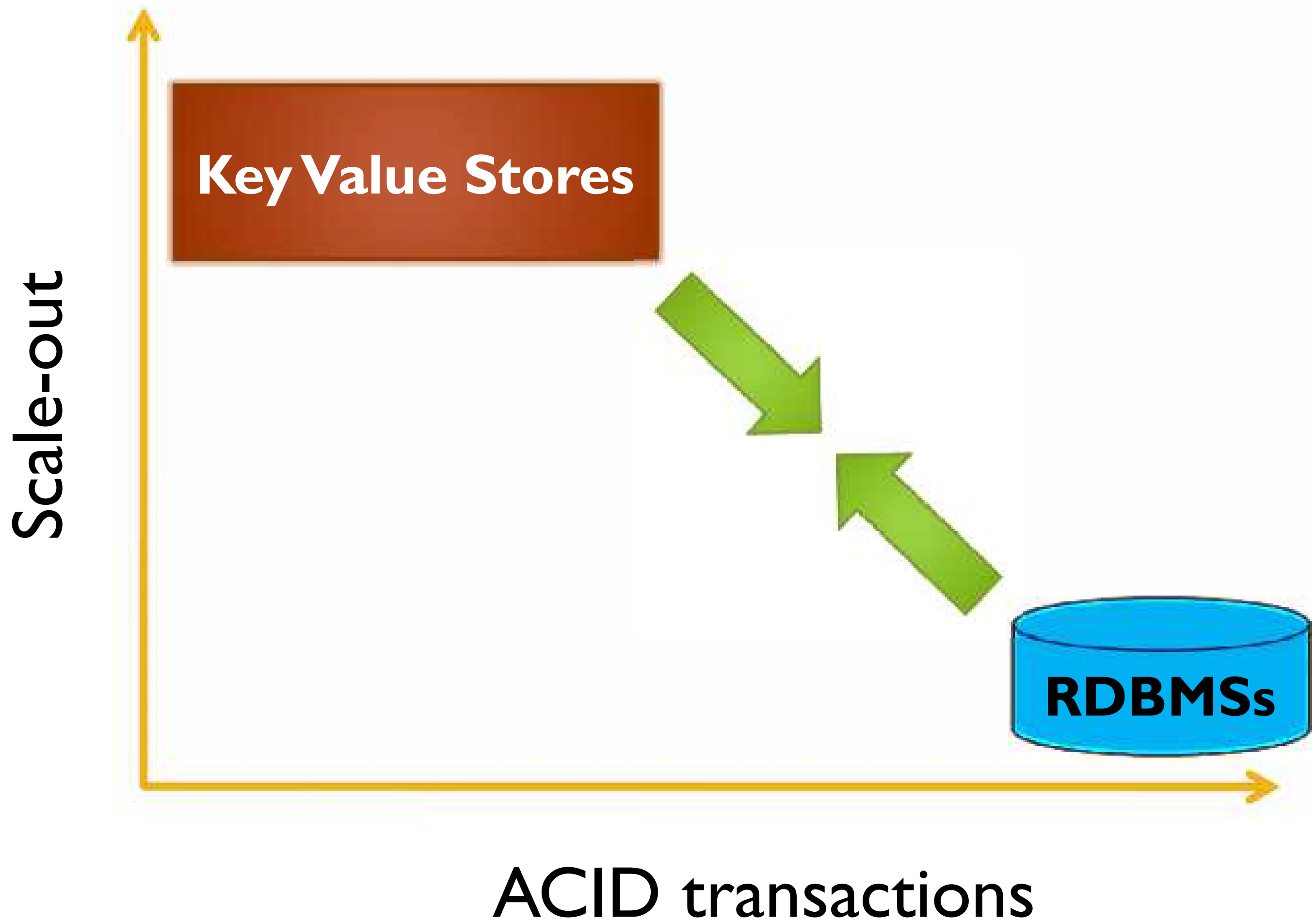
Thanks: Curino et al VLDB 2010

Design Principles

- **Limited distributed synchronization is practical**
  - Maintenance of metadata
  - Provide strong guarantees only for data that needs it



Challenge: Transactions at Scale



Transactions on Co-located Data

- Data or Ownership Co-location
  - Static partitioning
    - Leveraging schema patterns
    - Graph-based partitioning techniques
  - Application-specified dynamic partitioning
- Transaction Execution
- Data storage
  - Coupled storage
  - Decoupled storage
- Replication
  - Explicit Replication
  - Implicit Replication

Data or Ownership Co-location

- Co-located ownership or data frequently accessed together within a transaction
  - Minimize distributed synchronization
- Two design patterns
  - Static partitioning
    - Statically partition data based on schema patterns of applications' access patterns
  - Dynamic partitioning
    - Leverage application hints

Static Partitioning

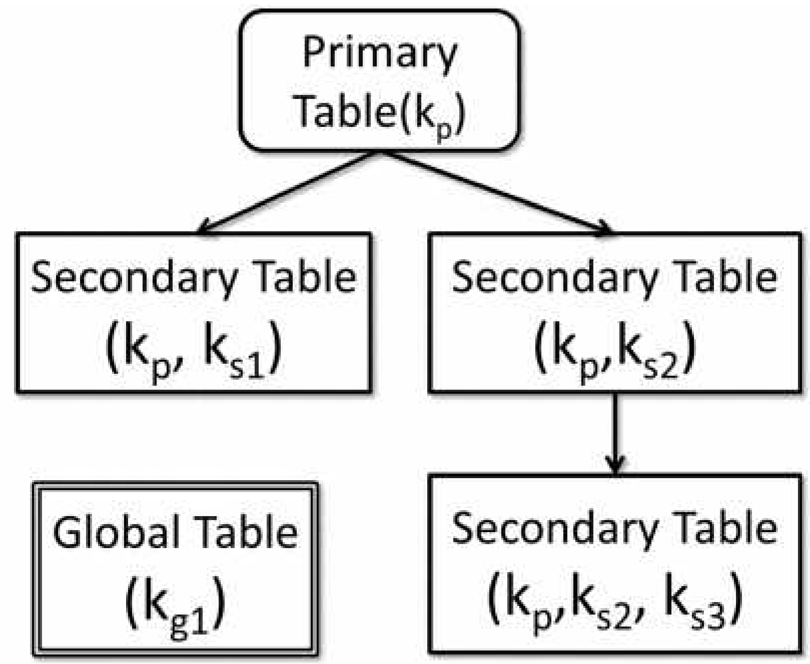
- Identify common schema design patterns across a class of applications
- Design applications conforming to these patterns by limiting data items accessed within a transaction
- Statically define the granule of transactional access
- Co-locate data (or ownership) of this granule

Leveraging Schema Patterns

- Hierarchy of objects or tables
- Transactions access data items forming this hierarchy
- Three common variants explored in literature
    - Tree Schema
    - Entity Groups
    - Table Groups

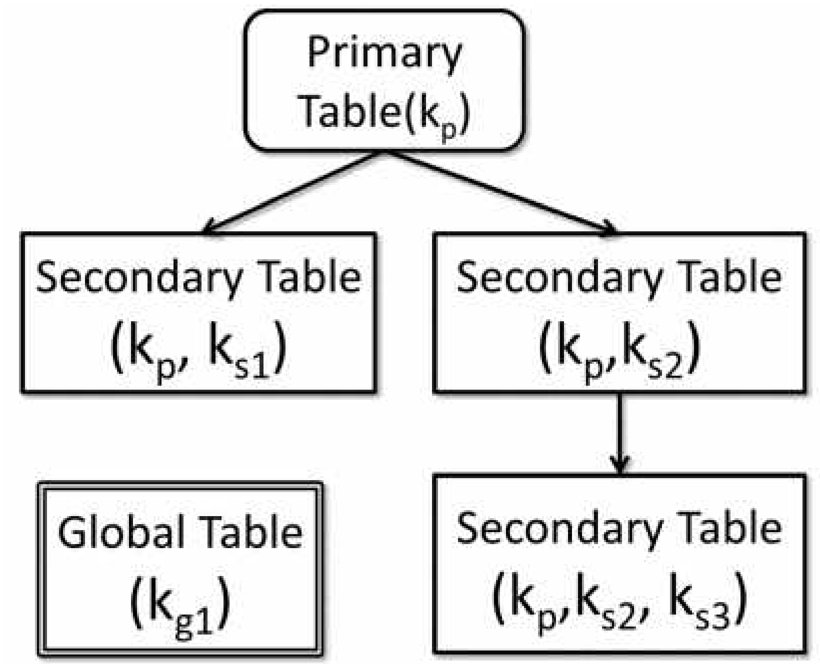
Tree Schema

- Tree types of tables
  - Primary
  - Secondary
  - Global
- Primary forms the root of the tree
  - Only one primary per schema
  - Key of primary acts as partitioning key



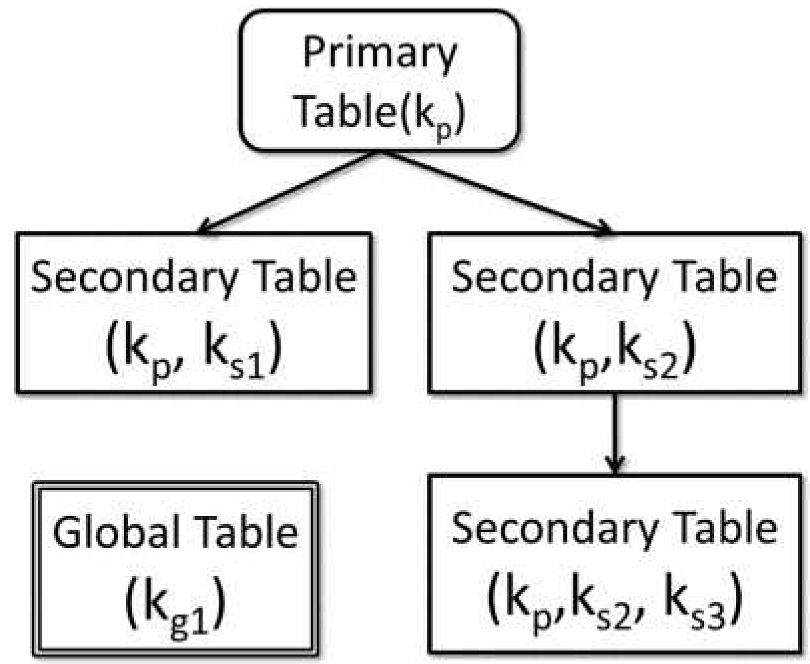
Tree Schema

- Secondary table have primary table's key as foreign key
- Global tables are lookup tables that are mostly read-only
- A schema can have multiple secondary and global tables



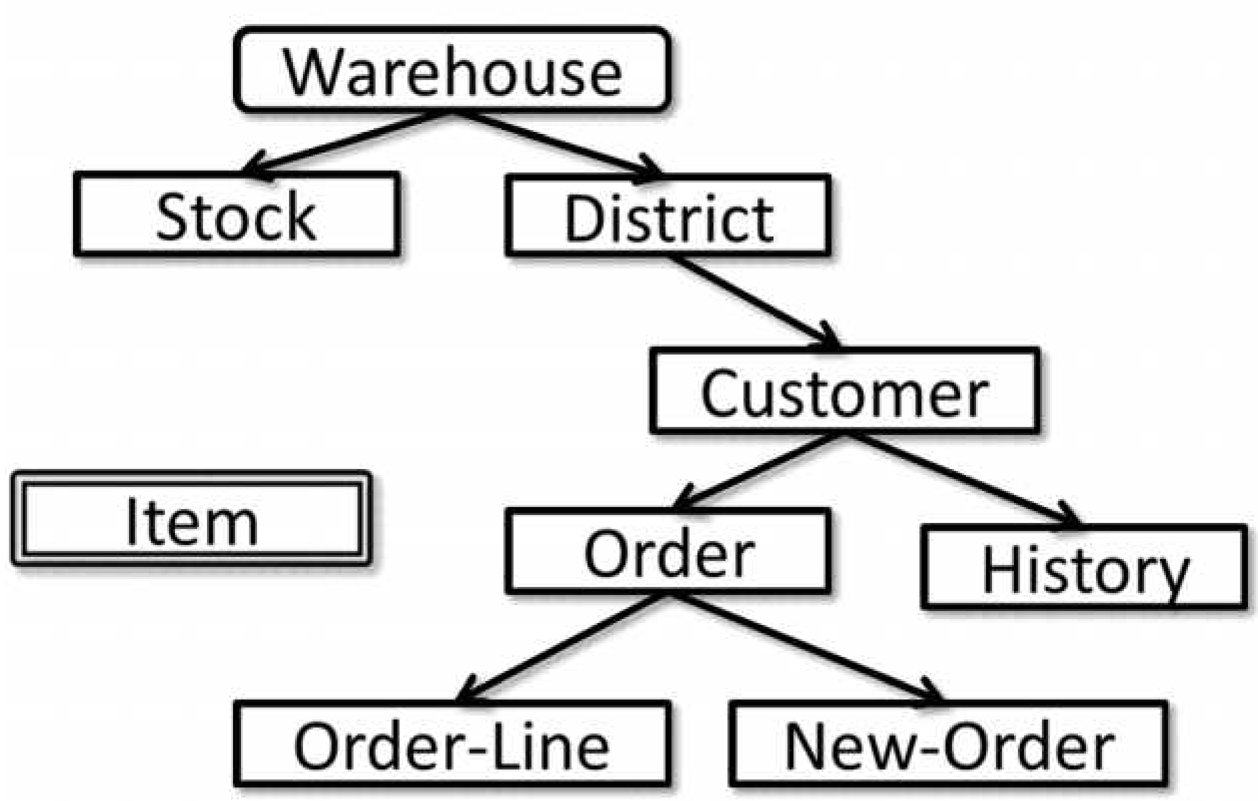
Tree Schema

- Corresponding to every row in the primary table, there are a group of related rows in the secondary tables
  - All rows referencing the same key form a *row group*
- All rows in a row group can be co-located
- Global tables replicated
- Transactions only access a row group or global tables



Tree Schema (example)

The TPC-C Schema is a tree schema



Entity Groups

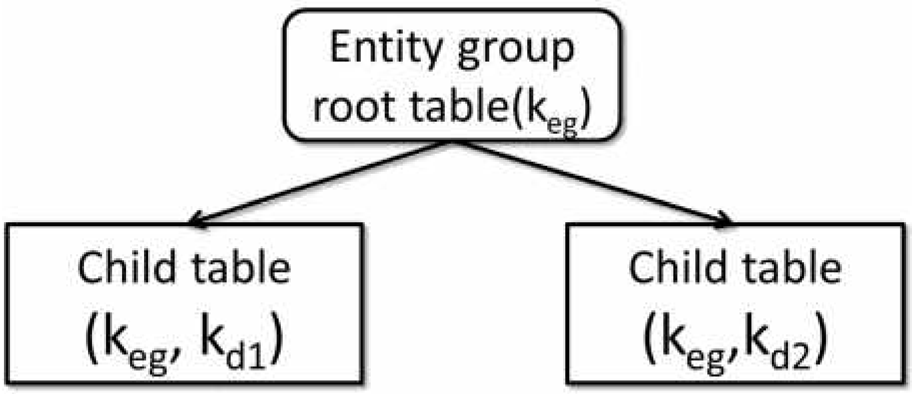
- Each schema consists or a set of tables
- Each table is comprised of a set of entities
- Each entity contains a set of properties which are named and typed columns
- Each table is either an *entity group root* table or a *child* table

Entity Groups

- Each child has a foreign key relationship with the root
- Each child entity refers to exactly one root entity
- A root entity along with all child entities that reference it form an entity group

➔ All entities forming a group can be co-located for efficient transactional access

Entity Groups



Entity group schema pattern

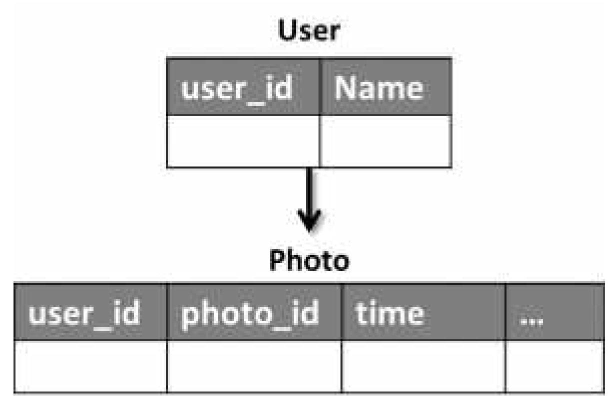


Photo App example

Table Groups

- A generalization of the tree schema and entity groups
- Table group consists of a set of tables
- *Keyed* table group
  - Partition key similar to entity groups/tree schema
  - All tables have column named partition key
  - Partition key need not be unique
  - All rows with same partition key form a *row group*
  - A partition is a set of row groups
- *Keyless* table group
  - More amorphous and more general

Discussion

- Hierarchical schema patterns allow data co-location
- Limit most, if not all, interactions to a single partition
- Multi-partition transactions can still be supported, but with higher cost
  - Distributed transactions
- Define a small unit of data as granule for consistent and transactional data access
  - Tight coupling within granules and loose coupling across granules

Access-driven Database Partitioning

- Analyze applications' access patterns
- Identify data items which, when co-located within a partition, will limit most transactions to a single partition
- Partition an application's data by analyzing its workload

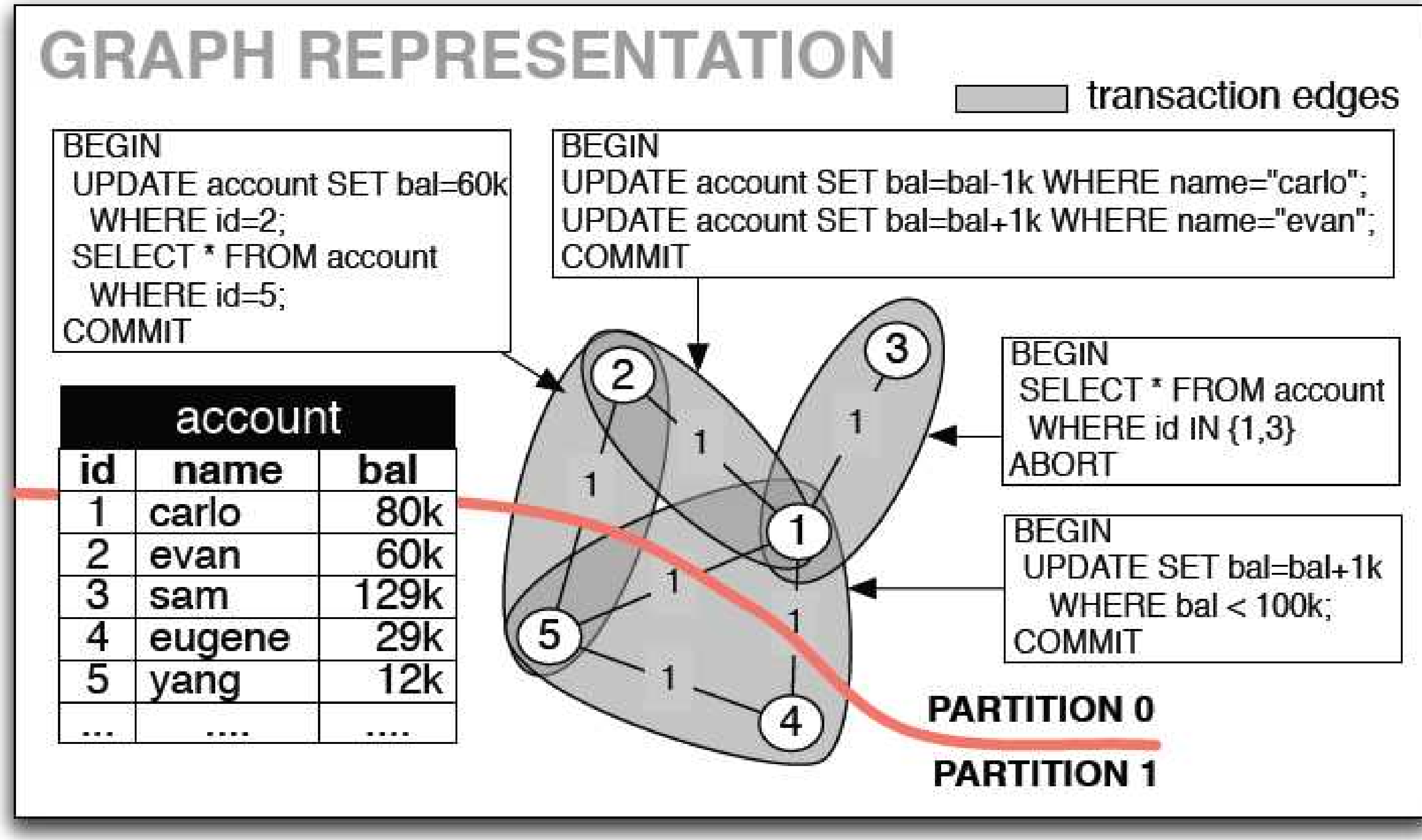
Schism – Graph-based partitioning

- A graph-based, data driven partitioning system for transactional workloads
- A database and its workload → using a graph, where tuples → nodes and transactions → edges connecting the tuples
- Partitioning the graph → minimum-cut partitioning of the graph into k partitions.

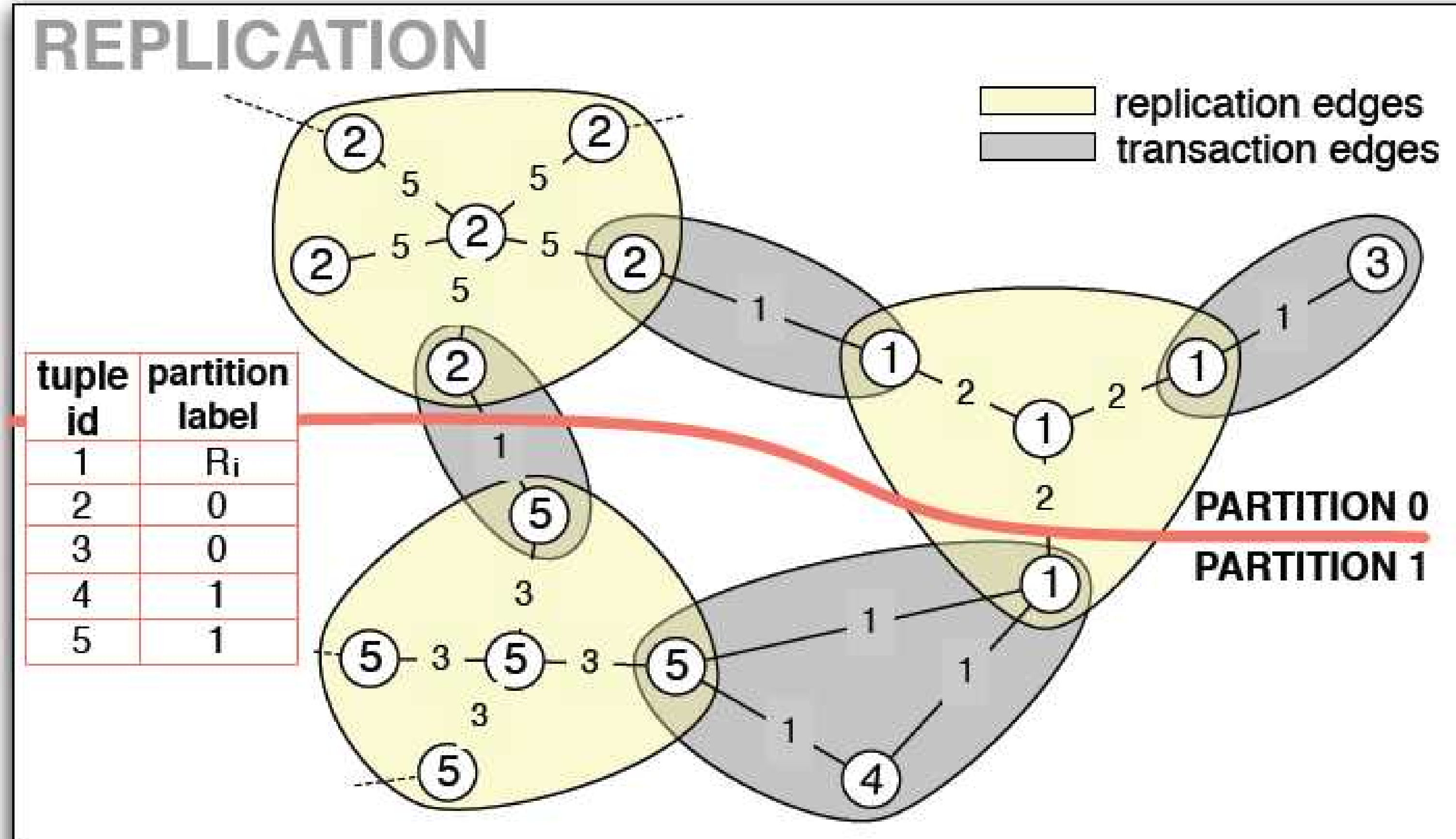
Overview of the Partitioning Algorithm

- Data pre-processing
  - Input transaction traces
  - Read and write sets
- Modeling interactions as a graph
  - Tuples (or rows) form nodes; accesses are edges
- Partitioning the graph
  - Balanced min-cut partitioning
- Explaining the partitioning
  - Learn a decision tree on frequent set of attributes

Graph Representation



Factoring in Replication

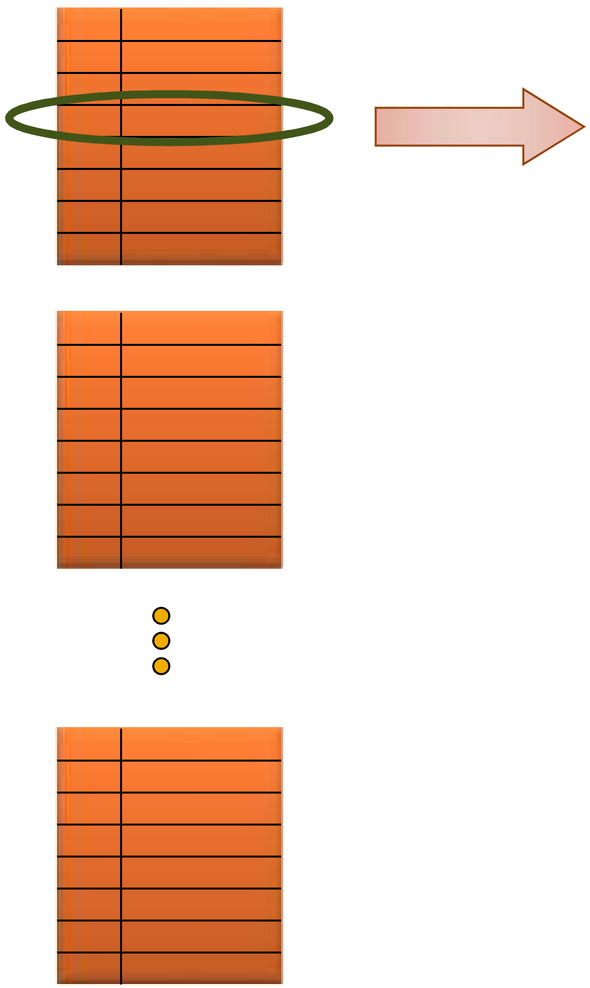


tuple id	partition label
1	R_i
2	0
3	0
4	1
5	1

What if partitions are not static?

- Access patterns change, often rapidly
  - Online multi-player gaming applications
  - Collaboration based applications
- Not amenable to static partitioning
- How to **co-locate transaction execution** when **accesses do not** statically **partition?**

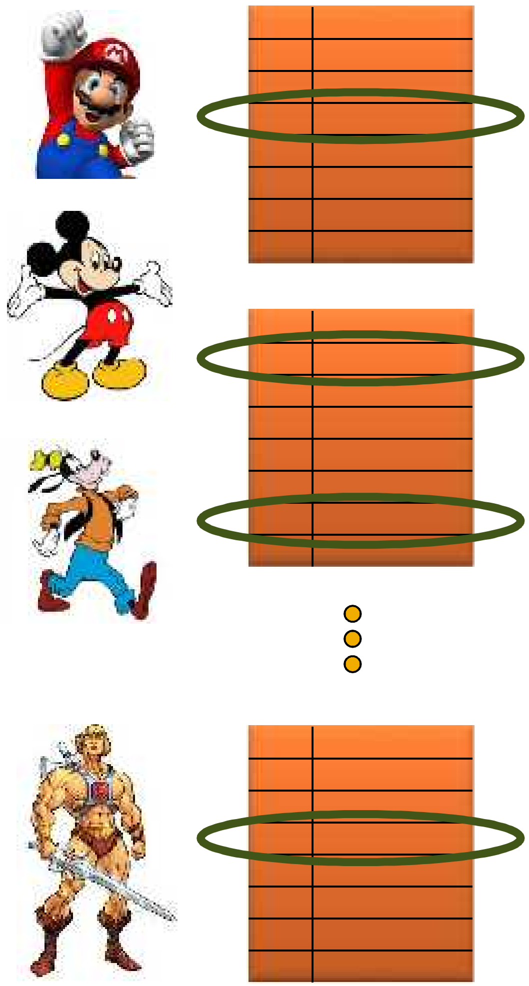
Online multi-player games



ID	Name	$$$	Score

Player Profile

Transactional access to a game

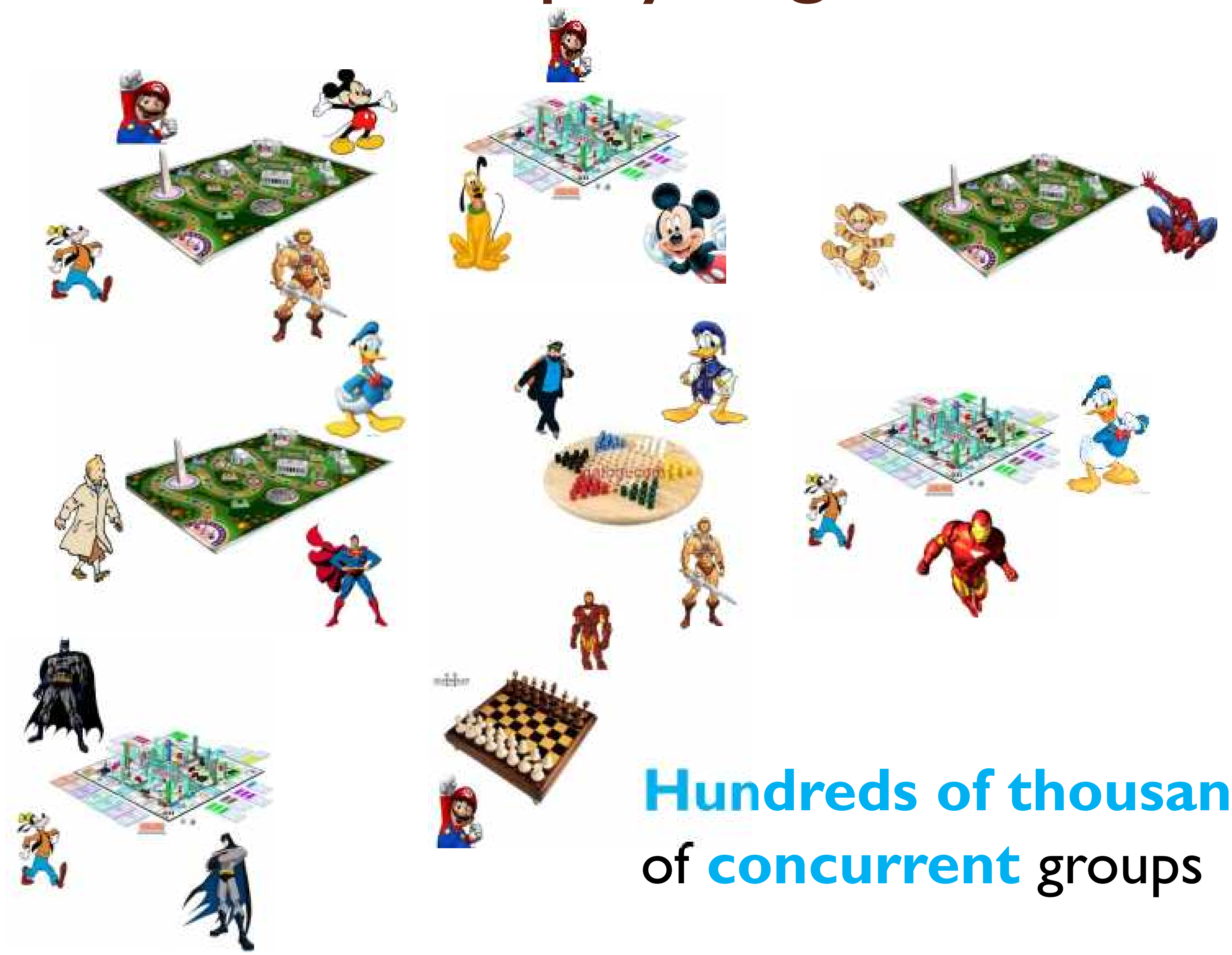


Execute transactions on player profiles while the **game is in progress**

Dynamics of gaming

Partitions/groups are **dynamic**

Scale of multi-player games

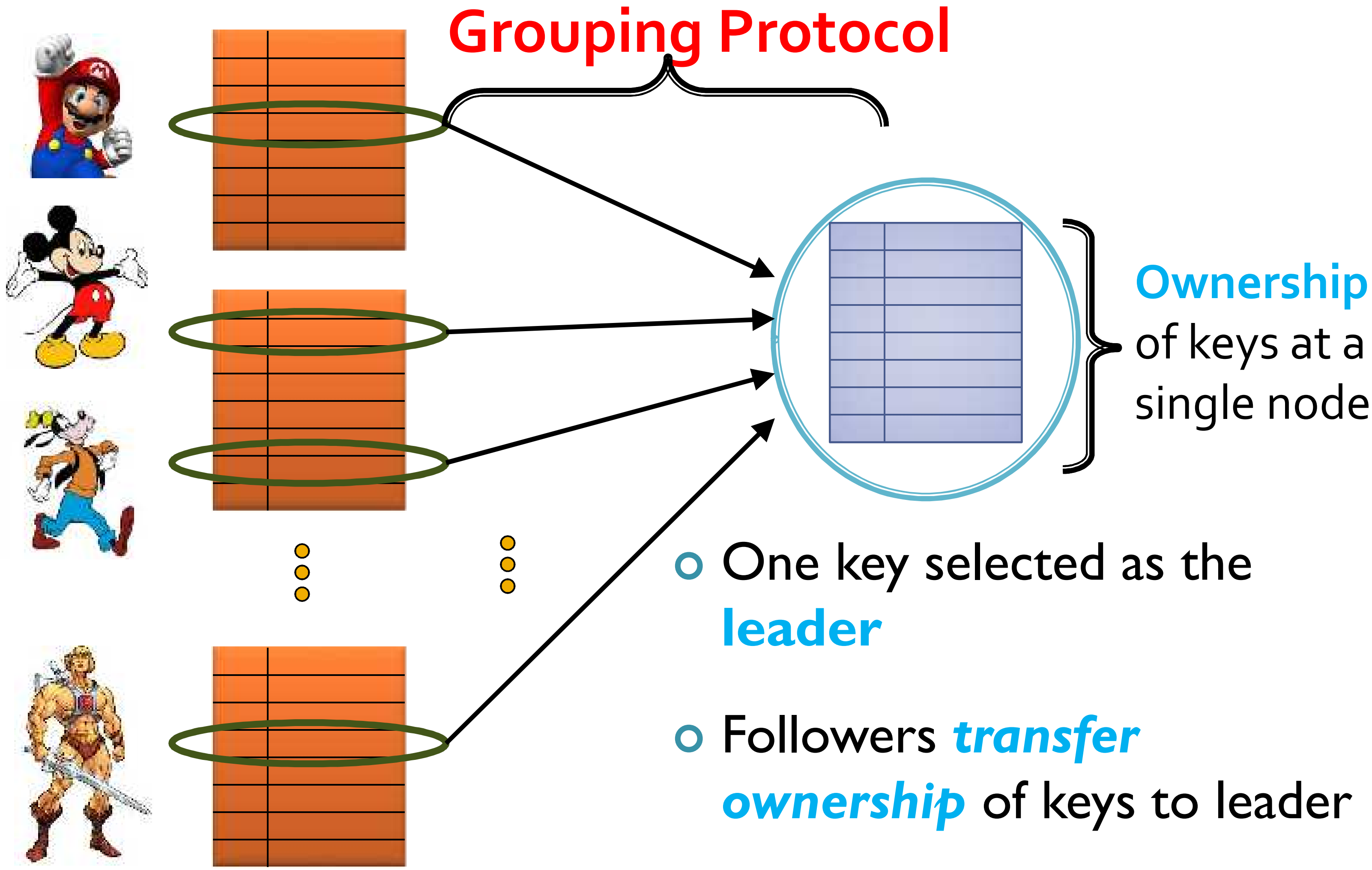


Hundreds of thousands of **concurrent** groups

The Key Group abstraction

- Allow applications to ***dynamically specify*** a ***group of data items***
- Support ***transactional*** access to the ***group*** formed ***on-demand***
- ***Challenge:*** Avoid distributed transactions!
- Properties of key groups
  - ***Small*** number of data items
  - Execute ***non-trivial no. of transactions***
  - ***Dynamic*** and ***on-demand***

Players in a game form a group



- One key selected as the **leader**
- Followers ***transfer ownership*** of keys to leader

Summary

- Techniques to co-locate data and/or ownership to limit transactions to a single node
- Hierarchical schema patterns
  - Tree schema
  - Entity groups
  - Table groups
- Access-driven partitioning
- Application-specified dynamic partitioning

Transaction Execution

- Once ownership is co-located, classical transaction processing techniques can be used
  - Leverage decades of research on concurrency control and recovery
- Concurrency control
  - Lock-based techniques
  - Optimistic concurrency control
- Recovery
  - Relies on logging. UNDO and/or REDO logging

Data Storage

- Conceptually, efficient non-distributed transaction execution only needs co-located ownership
- Two alternatives for physical data storage
  - Coupled Storage
  - Decoupled Storage

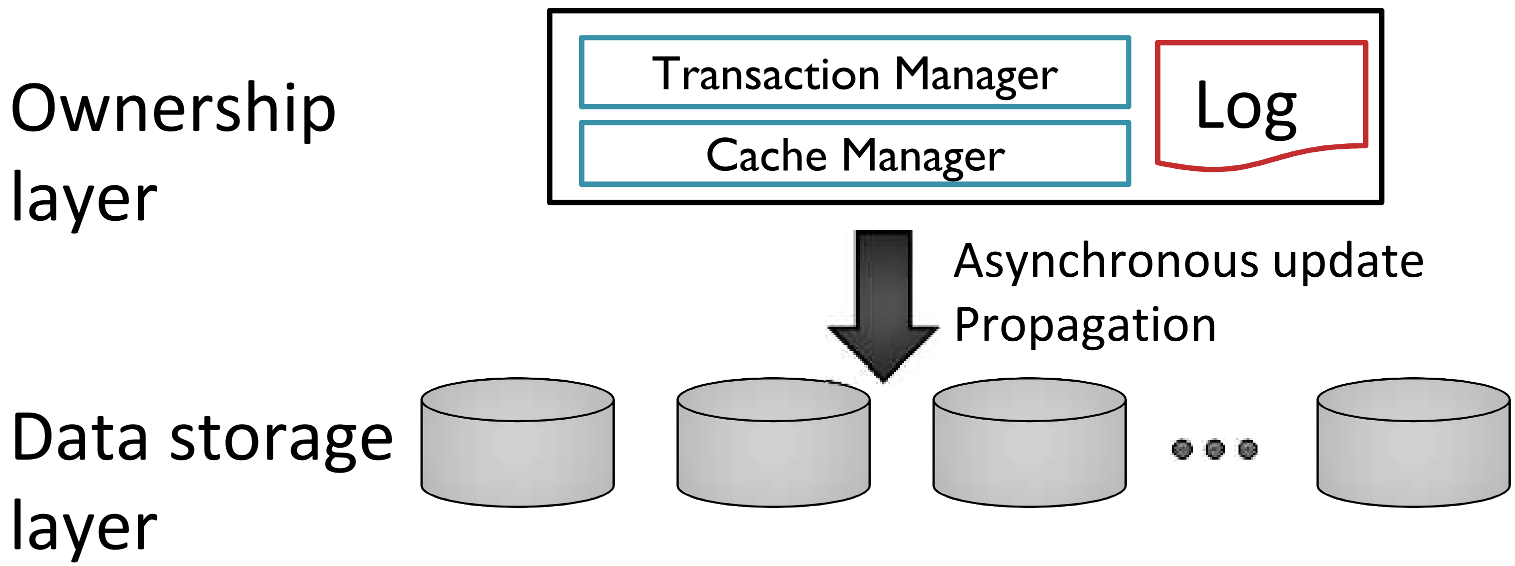
Coupled Storage

- Coupling storage with computation is a classical design choice for data intensive systems
  - Popularly known as the shared-nothing architecture
  - Improves performance by eliminating network transfers
- Side effect of co-locating ownership is that data items of a partition are also physically co-located

Decoupled Storage

- Data ownership is decoupled from the physical storage of data
  - Enablers: Growing main memory sizes and low latency high throughput data center networks
  - Rationale:
    - Working set typically fits in main memory for most OLTP systems
    - Large main memories can fit even larger working sets
    - Fast networks allow quick access to infrequent cache misses

Decoupled Storage Architecture



Benefits of Decoupled Storage

- Results in simplified design
  - Allows the storage layer to focus on fault-tolerance
  - Ownership layer can provide higher-level guarantees
- Allows independent scaling of the ownership and data storage layer
- Allows lightweight migration of ownership for elastic scaling and load balancing

Decoupled Storage Architectures

- Two alternative approaches explored in literature
  - Managed storage layer
    - Transaction manager controls the physical layout and format
    - The storage layer exposes an abstraction of a distributed replicated block storage
  - Self-managed storage layer
    - Transaction layer oblivious of physical design and layout

Managed Storage Layer

- Treat the decoupled storage layer as a distributed and replicated block storage
  - Examples: Google File System, Hadoop Distributed File System, Windows Azure Storage, Amazon S3
- Divide design complexity between transaction management and storage layers
- Transaction management layer: concurrency control, recovery, storage layout
- Storage layer: replication, geo-distribution, fault-tolerance, load balancing

Self-managed Storage Layer

- Provides more autonomy to the storage layer
- Transaction layer oblivious of the physical data layout and structures
- Transaction layer operates at the granularity of logical objects
- Hence can span different storage formats
  - B-trees, RDFS, Graph Stores

Replication

- The way a system handles data replication adds another dimension to the design space
- Synchronous or asynchronous replication
- Primary copy, multi-master, or quorum consensus
- Trade-offs related to consistency, availability, partition tolerance, performance, data durability, and disaster recovery
- Our focus: Explicit or Implicit Replication

Explicit Replication

- Design the transaction manager to be cognizant of replication
- Updates made by the transactions are explicitly replicated by the transaction manager
- Example: Primary-based replication in Cloud SQL Server and Megastore (inter-data center replication)
- Benefits: Quick failover from primary to secondary

Implicit Replication

- Data replication transparent to transaction execution
  - Typical in decoupled storage architectures
  - Storage layer manages replication
- Examples: ElasTraS, G-Store, Megastore (for intra-data center replication)
- Physical or logical replication

Summary

- Data or Ownership Co-location
  - Static partitioning
    - Leveraging schema patterns
    - Graph-based partitioning techniques
  - Application-specified dynamic partitioning
- Transaction Execution
- Data storage
  - Coupled storage
  - Decoupled storage
- Replication
  - Explicit Replication
  - Implicit Replication

References

- Alsberg, P.A., and Day, J. D.: A principle for resilient sharing of distributed resources. In ICSE, pp. 562-570, 1976.
- Attar, R., Bernstein, P.A, and Nathan Goodman: Site Initialization, Recovery, and Backup in a Distributed Database System. *IEEE Trans. Soft. Eng.* 10(6), pp. 645-650, 1984.
- Bailis, P., Venkataraman, S., Franklin, M. J., Hellerstein, J. M., and Stoica, I.: Probabilistically Bounded Staleness for Practical Partial Quorums. *PVLDB* 5(8), pp. 776-787, 2012.
- Bernstein, P.A. and Newcomer, E.: Principles of Transaction Processing, *Morgan Kaufmann*, 2nd ed., 2009.
- Brewer, E. A.: Towards Robust Distributed Systems (abstract). In PODC, p. 7, 2000.
- Burckhardt, S., Leijen, D., Fähndrich, M.,Sagiv, M.: Eventually Consistent Transactions, In ESOP, pp. 67-86, 2012.
- Cooper, B. F., Ramakrishnan, R., Srivastava, U., Silberstein, A., Bohannon, P., Jacobsen, H., Puz, N., Weaver, D., and Yerneni., R.: PNUTS: Yahoo!'s hosted data serving platform. *Proc. VLDB Endow.* 1(2), pp. 1277-1288, 2008.
- Davidson, S. B., Garcia-Molina, H. and Skeen, D.: Consis-tency in a Partitioned Network: a Survey. *ACM Comput. Surv.* 17(3), pp. 341-370, 1985.
- DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., and Vogels, W.: Dynamo: Amazon's highly available key-value store. In SOSP, pp. 205-220, 2007.
- Ellis, C. A. and Gibbs, S. J.: Concurrency control in Groupware systems. In SIGMOD, pp. 399-407, 1989.
- Fischer, M. J. and Michael, A.: Sacrificing Serializability to Attain High Availability of Data in an Unreliable Network. In PODS, pp. 70-75, 1982.

References

- Gilbert, S. and Lynch, N. A.: Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *SIGACT News* 33(2), pp. 51-59, 2002.
- Herlihy, M. P. and Wing, J. M.: Linearizability: a Correctness Condition for Concurrent Objects. *ACM Trans. Program. Lang. Syst.* 12(3), pp. 463-492, 1990.
- Kemme, B. and Alonso, G.: Database Replication: a Tale of Research across Communities. *PVLDB*, 3(1), pp. 5-12, 2010.
- Ladin, R., Liskov, B., Shrira, L., and Ghemawat, S.: Providing high availability using lazy replication. *ACM Trans. Comput. Syst.* 10 (4), 360-391, 1992.
- Lamport, L.: Time, clocks, and the ordering of events in a distributed system. *Comm. ACM,* 21(7), pp. 558-565, 1978.
- Li, C., Porto, D., Clement, A., Gehrke, J., Preguic, N., and Rodrigues, R.: Making Geo-Replicated Systems Fast if Possible, Consistent when Necessary. OSDI, pp. 265-278, 2012.
- Lloyd, W., Freedman, M. J., Kaminsky, M., and Andersen, D. G.: Don't settle for eventual: scalable causal consistency for wide-area storage with COPS. SOSP, pp. 401-416, 2011.
- Lloyd, W., Freedman, M.J., Kaminsky, M. and Andersen, D.G.: Stronger Semantics for Low-Latency Geo-Replicated Storage. NSDI '13, pp. 313-328, 2013.
- Parker Jr., D. S., Popek, G. J., Rudisin, G., Stoughton, A., Walker, B.J., Walton, E., Chow, J.M., Edwards, D., Kiser, S., Kline, C. : Detection of Mutual Inconsistency in Distributed Systems. *IEEE Trans. Software Eng* 9(3), pp. 240-247, 1983.
- K. Petersen, M. Spreitzer, D. Terry, M. Theimer, and A. Demers. Flexible update propagation for weakly consistent replication. In SOSP, Oct. 1997.
- Rothnie. J. B., and Goodman, N.: A Survey of Research and Development in Distributed Database Management. In VLDB, pp. 48-62, 1977.
- Saito, Y. and Shapiro, M.: Optimistic replication. *ACM Comput. Surv.* 37(1), pp. 42-81, 2005.
- Shapiro, M., Preguiça, N., Baquero, C., Zawirski, M.: Convergent and Commutative Replicated Data Types. *Bulletin of the EATCS*, No.104, pp. 67-88, 2011.
- Stonebraker, M. and Neuhold, E. J.: A Distributed Database Version of INGRES. Berkeley Workshop, pp. 19-36, 1977.
- Sovran, Y., Power, R., Aguilera, M. K., and Li, J.: Transactional Storage for Geo-replicated Systems. In SOSP, pp. 385-400, 2011.
- Terry, D. B.: Replicated Data Management for Mobile Computing. *Morgan Claypool Publishers*, 2008.
- Terry, D.B.: Replicated Data Consistency Explained Through Baseball, MSR-TR-2011-137, http://research.microsoft.com
- Terry, D. B., Demers, A. J., Petersen, K., Spreitzer, M.J., Theimer, M.M., Welch, B. B.: Session guarantees for Weakly Consistent Replicated Data. In PDIS, pp. 140-149, 1994.
- Thomas, R. H.: A majority consensus approach to concurrency control for multiple copy databases. *ACM Trans. on Database Systems,* 4(2), pp. 180–209, 1979.