THE UNIVERSITY *of* EDINBURGH

Distributed Systems Fall 2024

Yuvraj Patel

Disclaimer: Slides prepared using multiple sources (UW-Madison – Remzi, Andrea, Mike; Cambridge – Martin Kleepman; Distributed Systems book by MVS/ AST); University of Edinburgh – Yuvraj Patel

Today's Agenda

Replication

- General Stuff
- Data-Centric Consistency Models
- Client-Centric Consistency Models

Replication

Replicate data at one or more sites can help with

- Availability & Fault Tolerance
  - If primary server crashes, secondary can takeover => Highly available service
  - Mask node crashes => Transparency
- Performance
  - Local access faster than remote access; Low latency
  - Concurrent Reads can be served from multiple servers improving performance
- Scaling
  - Size scalability – Prevent overloading a single server
  - Geographical scalability

Problems with Replication

Having multiple copies, means that when any copy changes, the change needs to be propagated to all other copies

Need replicas to have same data, i.e., they should be kept consistent

Efficiently synchronize all replicas a challenging problem

Performance & Scalability

Main concern – To keep replicas consistent, we generally need to ensure that all conflicting operations are done in the same order, across all servers

Conflicting operations – Recall the read-write and write-write conflicts

Guaranteeing global ordering on conflicting operations may be costly operation, with impact on scalability

Potential Solution – Weaker consistency requirements to avoid global synchronization, whenever possible

Weakening Consistency Requirements

What does it mean to weaken consistency requirements?

- Relax the requirement that “updates need to be executed as atomic operations”
- Do not require global synchronizations
- Replicas may not always be the same everywhere and everytime

To what extent can consistency be weakened?

- Depends highly on the access and update patterns of the replicas
- Depends on the replicated data user patterns which is application driven

Consistency Models

A consistency model is a contract between the programmer and a system

- The system guarantees that if the programmer follows the rules for operations on data, data will be consistent
- Result of the reading, writing, updating data will be predictable

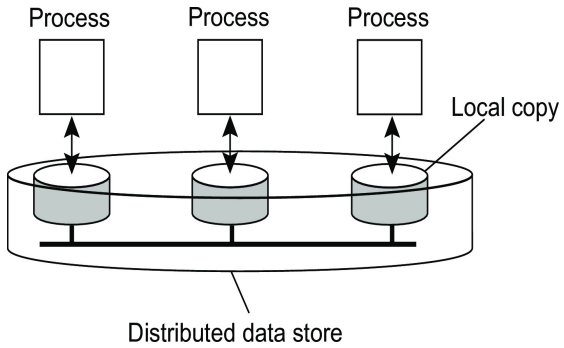
Two consistency models

- Data-centric consistency models – Defines consistency as experienced by all the clients; provides a system wide consistent view on the data store
- Client-centric consistency models – Defines consistency of the data store only from one client's perspective; Different clients might see different sequences of operations at their replicas

Distributed Data Store

Distributed Data Store – Physically distributed & replicated across multiple machines

- Data can be read from or written by any process on any node
- A local copy helps with faster reads
- A write to a local replica needs to be propagated to all remote replicas

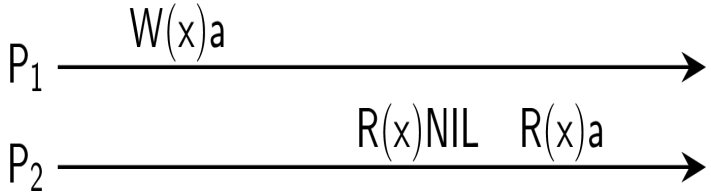


Terminology & Notations

Read and write operations

- $W_i(x)a$: Process P_i writes value a to x
- $R_i(x)b$: Process P_i reads value b from x
- All data items initially have value NIL

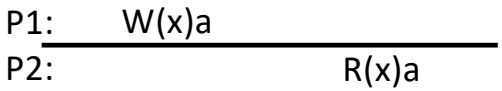
Possible behavior represented over time; time moves from left to right



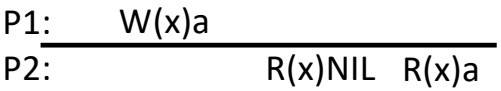
Strict Consistency

With strict consistency, all writes are visible instantaneously to all processes

Any read to a shared data item returns the value stored by the most recent write operation on that data item



Strictly Consistent Data Store



Not Strictly Consistent Data Store

Strictest consistency model – most rigid model

Practical relevance restricted to a thought experiment and formalism

- Relies on absolute global time
- Instantaneous message exchange is impossible

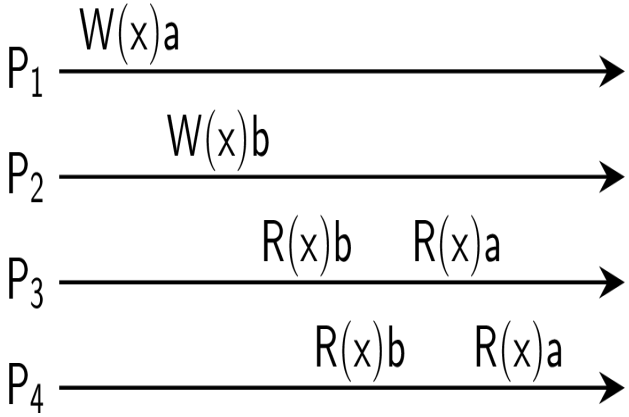
Sequential Consistency

Sequential Consistency – The result of any execution is the same as if the operations by all processes were executed in some sequential order and the operations of each individual process appear in this sequence in the order specified by its program

Any valid interleaving of read or write operations is fine, but all processes must see the same interleaving

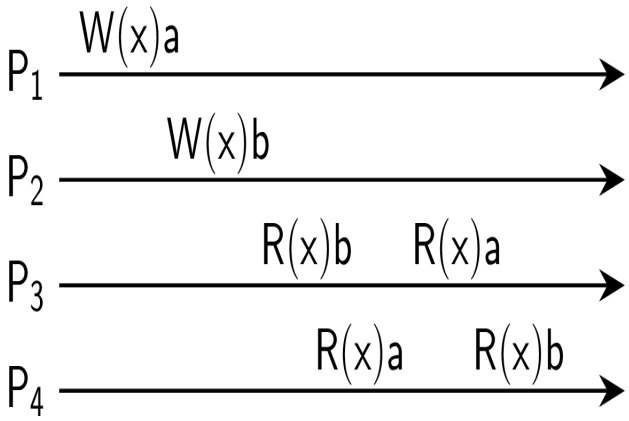
- The events observed by each process must globally occur in the same order, or it is not sequentially consistent

Sequential Consistency Example



A sequentially consistent data store

P3 and P4 see the same interleaving of writes



A data store that is not sequentially consistent

P3 and P4 do not see the same interleaving of writes

Linearizability

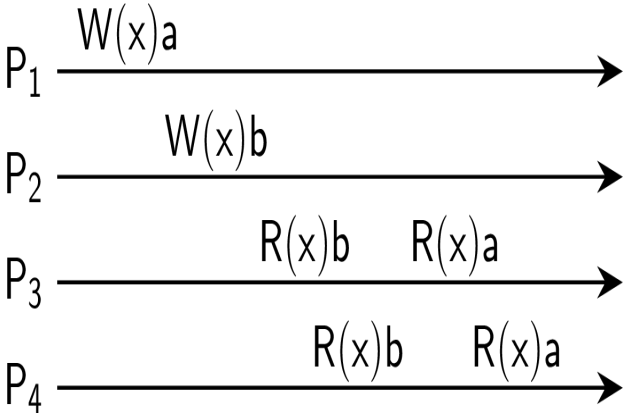
In sequential consistency, absolute time is somewhat irrelevant – the order of events is most important

Linearizability – Each operation should appear to take effect instantaneously at some moment between its start and completion

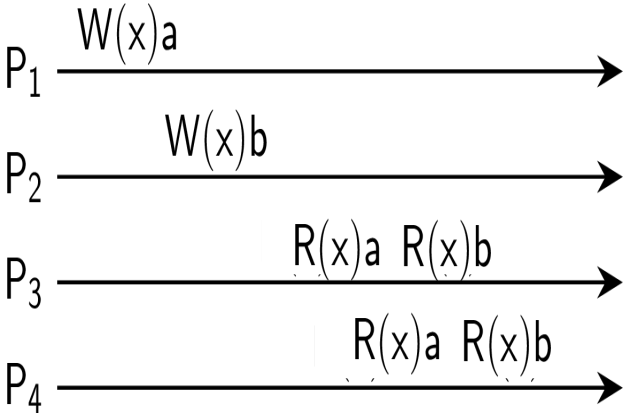
A data store is said to be linearizable when each operation is timestamped, and the following conditions hold:

- Sequential Consistency holds
- Timestamp($OP_1(x)$) < Timestamp($OP_2(x)$) then $OP_1(x)$ should precede $OP_2(x)$ in this sequence

Linearizability Consistency Example



Is this linearizable?



A linearizable consistent data store

Sequential Consistency vs. Linearizability

Linearizability is weaker than strict consistency, but stronger than sequential consistency

Linearizability cares about time; sequential consistency cares about program order

- With Sequential consistency, the system has freedom of how to interleave operations coming from different clients, as long as the ordering from each client is preserved
- With Linearizability, the interleaving across all clients is pretty much determined already based on the time

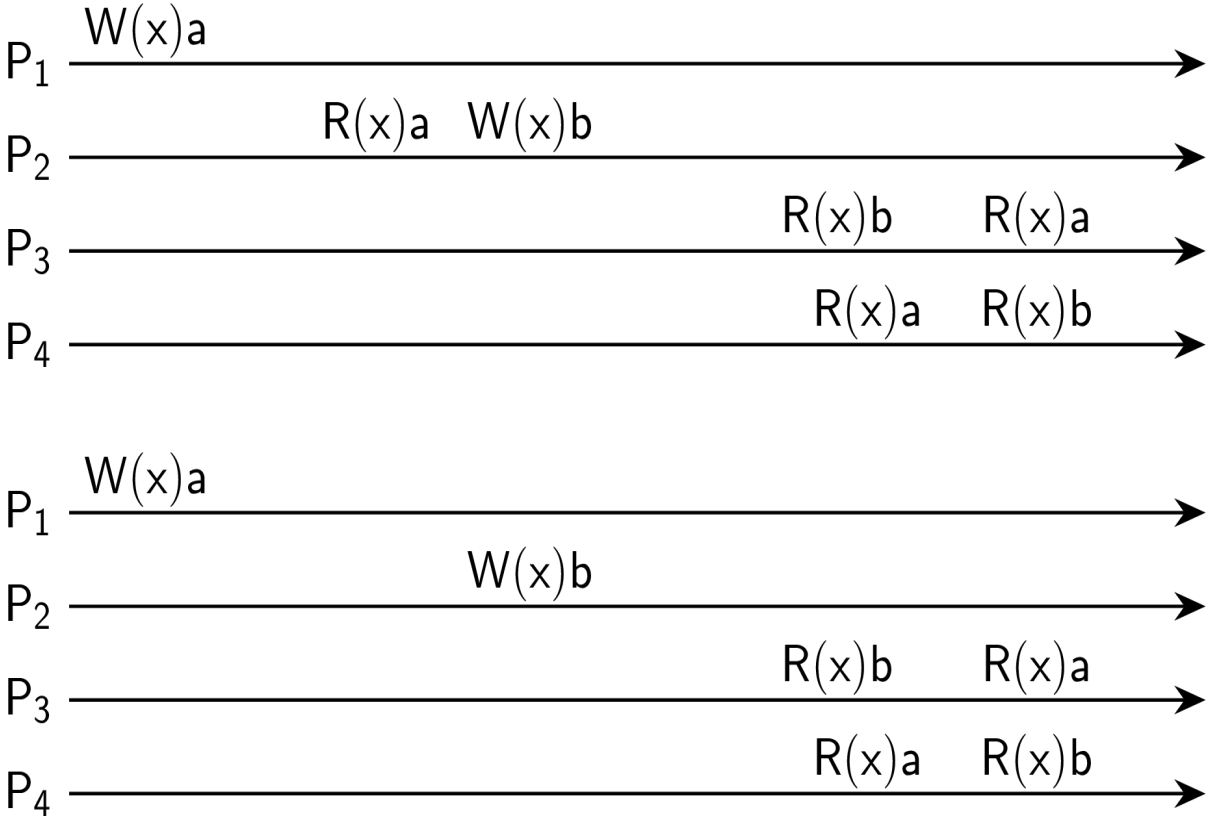
Causal Consistency

Writes that are potentially causally related must be seen by all processes in the same order

Concurrent writes may be seen in a different order on different machines

Example – If event B is a direct or indirect result of another prior event A, then all processes should observe event A before observing event B

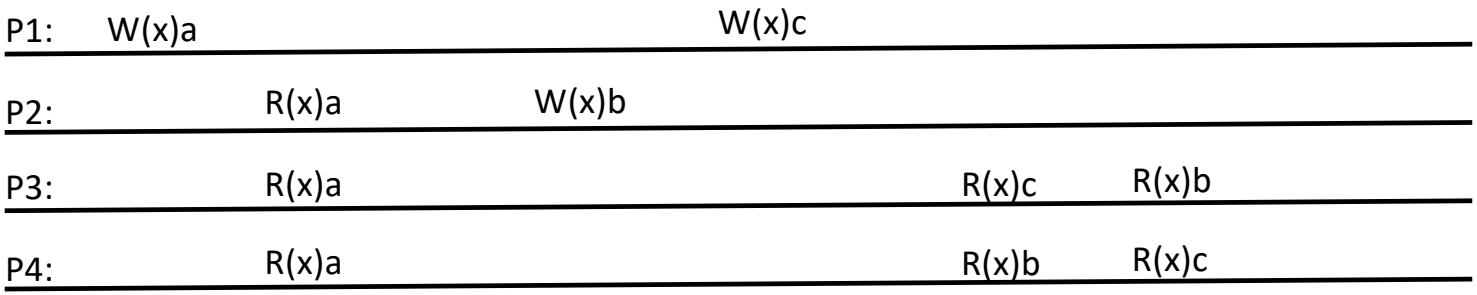
Causal Consistency Example



A violation of a causally-consistent store

A correct sequence of events in a causally-consistent store

Causal Consistency Example



Assume $W_2(x)b$ and $W_1(x)c$ are concurrent

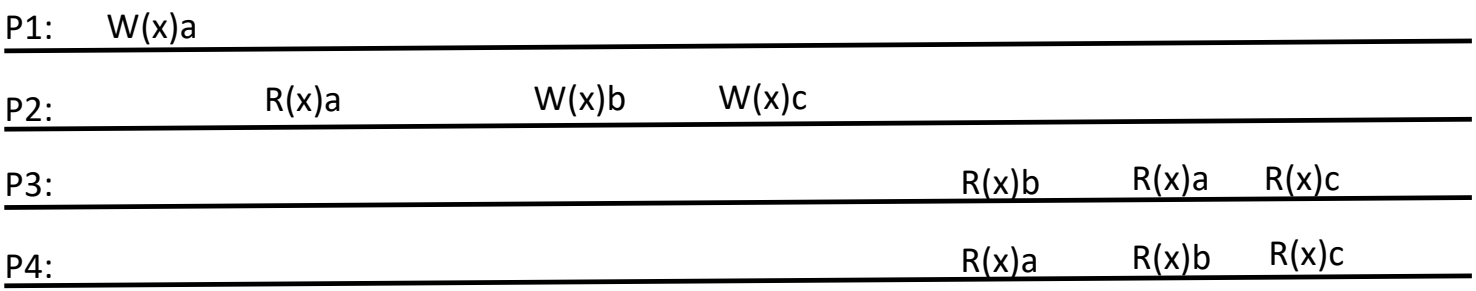
Strictly consistent?
Sequentially consistent?
Causally consistent?

FIFO Consistency

Writes performed by a single process are seen by all other processes in the order in which they were issued

Writes from different processes may be seen in a different order by different processes

FIFO consistency is easy to implement



A valid sequence of events of FIFO consistency
(P2's writes are seen in the correct order)

Data-Centric Consistency – Strong & Weak Models

Strong Consistency Models – Operations on shared data are synchronized; do not require synchronization operations

- Strict Consistency – Absolute time ordering of all shared accesses matters
- Sequential Consistency – All processes see all shared accesses in the same order
- Linearizability – Sequential Consistency + Operations are ordered according to a global time
- Causal Consistency – All processes see causally-related shared accesses in the same order
- FIFO Consistency – All processes see writes from each other in the order they were used

Weak Consistency Models – Synchronization occurs only when shared data is locked and unlocked; rely on synchronization operations

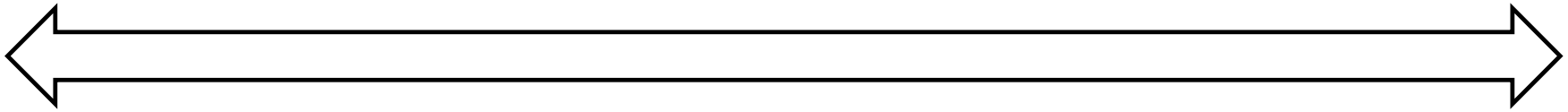
- Weak Consistency – Shared data can be counted on to be consistent only after a synchronization is done
- Release Consistency – Shared data are made consistent when a critical region is exited
- Entry Consistency – Shared data pertaining to a critical region are made consistent when a critical region is entered

Weaker the consistency models, the more scalable it is

Data Replication in Cloud

Client-Centric Consistency Models

Spectrum of Consistency



Lots of consistencies proposed in research community

Today, we will consider the six Guarantees discussed by Doug Terry

Baseball Game

One game comprises 9 innings

Game starts with 0-0 score

Visitors bat first and remain at bat until they make three outs

Then home team bats until they make three outs

Continue for 9 innings

Key-Value Store for Score

Score recorded in a Key-Value (KV) store in two objects

One object for visitor's score, another for home team

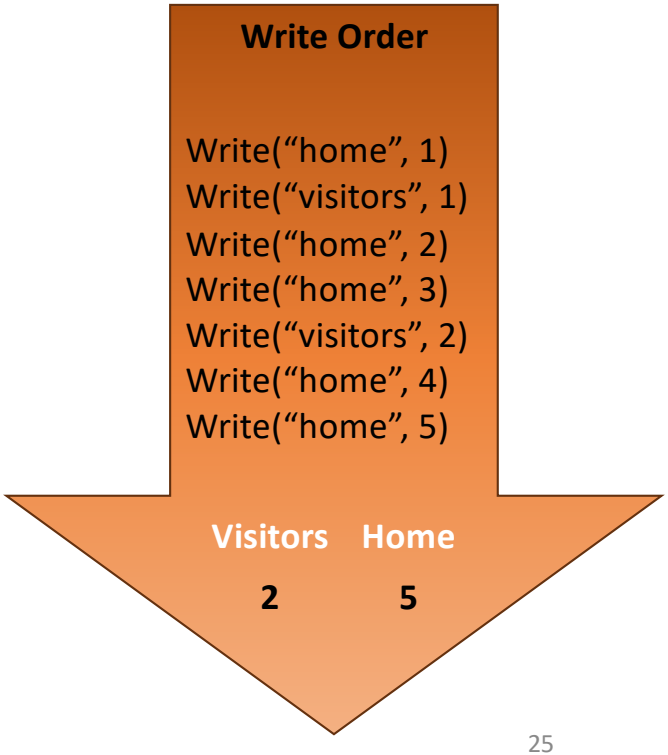
When a team scores a run,

- Read operation is performed on its current score
- The returned value is incremented by 1
- The new value is written back to the KV store

Sample Game Score

Existing Score

	1	2	3	4	5	6	7	8	9	RUNS
Visitors	0	0	1	0	1	0	0			2
Home	1	0	1	1	0	2				5



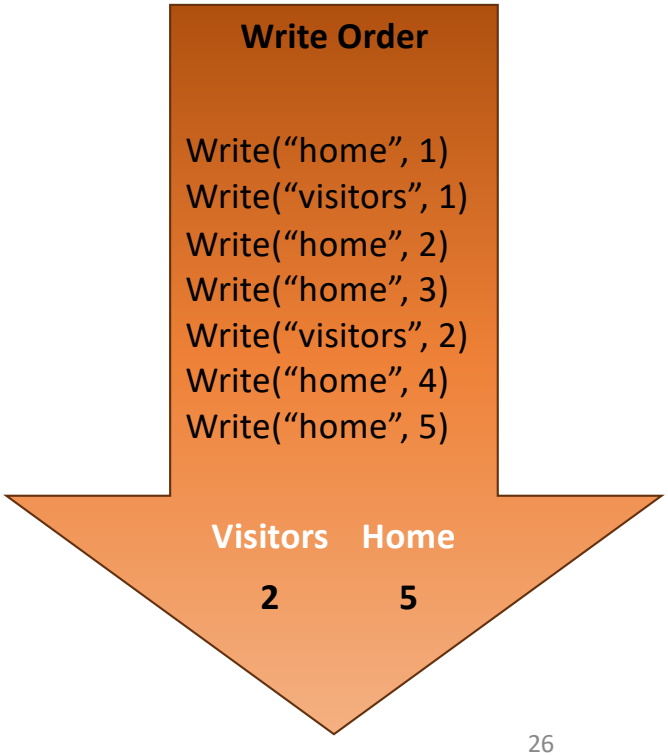
Strong Consistency

Guarantee – See all previous writes

All readers read the same data

Possible values

2, 5



Eventual Consistency

Guarantee – See subset of previous writes

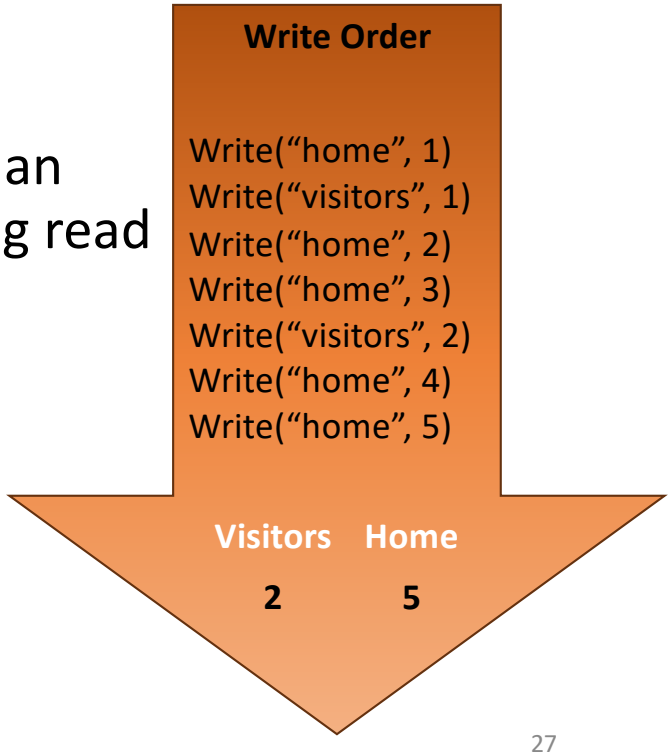
Readers can read data that is written in the past

Read can return from a replica that has received an arbitrary subset of writes to the data object being read

Eventually see all writes

Possible values

0-0, 0-1, 0-2, 0-3, 0-4, 0-5, 1-0, 1-1, 1-2, 1-3, 1-4, 1-5, 2-0, 2-1, 2-2, 2-3, 2-4, 2-5



Consistent Prefix

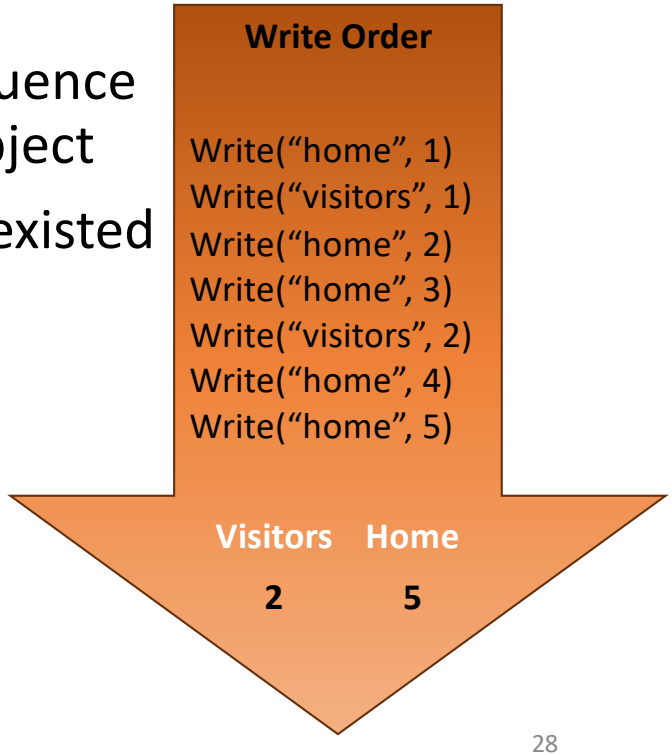
Guarantee – See initial sequence of writes

Reader is guaranteed to observe an ordered sequence of writes starting with the first write to a data object

The reader sees a version of the data store that existed at the primary at some time in the past

Possible values

0-0, 0-1, 1-1, 1-2, 1-3, 2-3, 2-4, 2-5



Bounded Staleness

Guarantee – See all ”old” writes

Reader ensured that read results are not too out-of-date

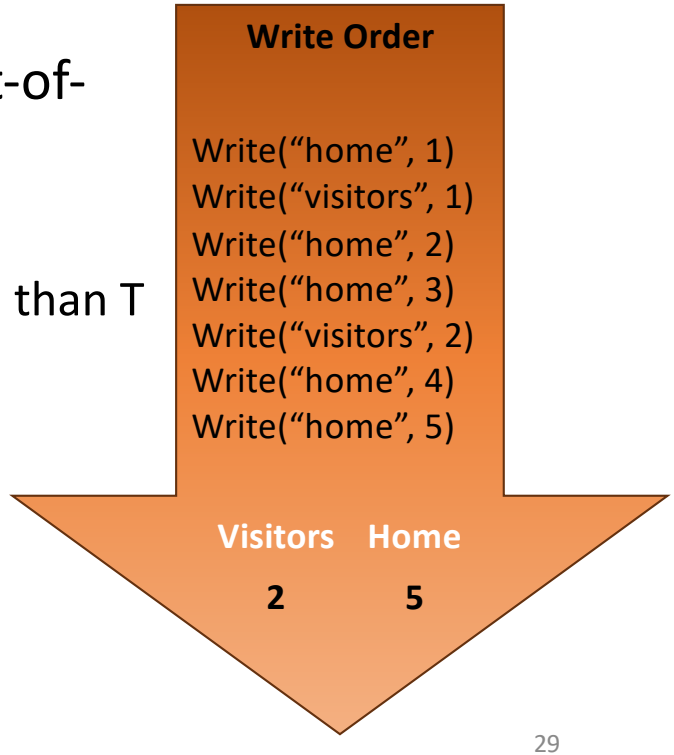
Staleness factor

- Read operation will return any values written more than T minutes ago or more recently written value

Possible values

At most one inning out-of-date score like

2-3, 2-4, 2-5



Monotonic Reads

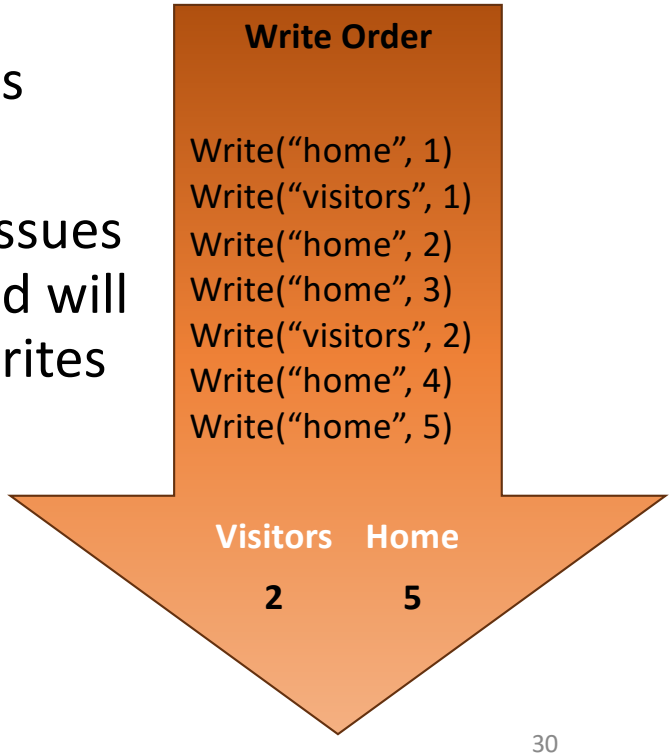
Guarantee – See increasing subset of writes

Reader guaranteed to observe a data store that is increasingly up-to-date over time

If reader issues a read operation and then later issues another read to the same object; the second read will return the same value(s) or the results of later writes

Possible value

After reading 1-3 – 1-3, 1-4, 1-5, 2-3, 2-4, 2-5



Read My Writes

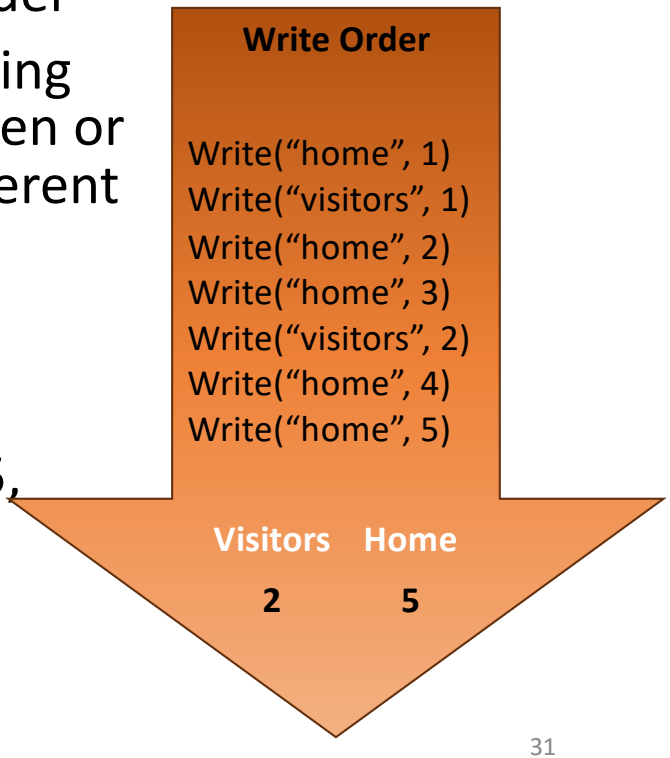
Guarantee – See all writes performed by the reader

After writing a new value and subsequently reading the value will return the value that was last written or some other value that was later written by a different client

Possible value

For the writer: 2-5

For anyone else: 0-0, 0-1, 0-2, 0-3, 0-4, 0-5, 1-0, 1-1, 1-2, 1-3, 1-4, 1-5, 2-0, 2-1, 2-2, 2-3, 2-4, 2-5



Summary

Different read guarantees

- Strong Consistency – See all previous writes
- Eventual Consistency – See subset of previous writes
- Consistent Prefix – See initial sequence of writes
- Bounded Staleness – See all “old” writes
- Monotonic Reads – See increasing subset of writes
- Read My Writes – See all writes performed by the reader

Consistency Trade-off

Trade-off between three properties – Consistency, Availability, Performance

Some entries may vary based on implementation, deployment, operating details, etc.

	Consistency	**Performance**	**Availability**
Strong Consistency	Excellent	Poor	Poor
Eventual Consistency	Poor	Excellent	Excellent
Consistent Prefix	Okay	Good	Excellent
Bounded Staleness	Good	Okay	Poor
Monotonic Reads	Okay	Good	Good
Read My Writes	Okay	Okay	Okay

Qualitatively Accurate General Comparison