

Transaction

Johan Montelius

KTH

HT15

Even if we have a distributed system that provides *atomic operations* we sometimes want to group a sequence of operations in a transaction where:

- either all are executed or
- none is executed
- even if a node crash

1 / 42

2 / 42

Recoverable objects: a server can store information in persistent memory (the file system) and can recover objects when restarted.

The service will not be *highly available*, but this is good enough for now.

3 / 42

4 / 42

- Atomic - either all or nothing
- Consistent - the server should be left in a consistent state
- Isolation - total order of transactions
- Durability - persistent, once acknowledged

All requirements can be achieved by only allowing sequential access to the transaction server.

Our goal is to provide as much concurrency as possible while preserving the behavior of sequential access.

What is the problem?

5 / 42

6 / 42

- `openTransaction()` : returns a transaction identifier (`tid`)
- `operation(tid, arg)` : the operations of the transaction
- `closeTransaction(tid)` : returns success or failure of transaction
- `abortTransaction(tid)` : client explicitly aborts transaction

We will write operations with implicit *tid*.

Operations:

- `getBalance()`
- `setBalance()`
- `withdraw(amount)`
- `deposit(amount)`

7 / 42

8 / 42

the lost update

client a

```
bal = b.getBalance()
↓
b.setBalance(bal*1.1)
↓
a.withdraw(bal*0.1)
```

client c

```
bal = b.getBalance()
↓
b.setBalance(bal*1.1)
↓
c.withdraw(bal*0.1)
```

9 / 42

inconsistent retrieval

client p

```
a.withdraw(100)
↓
b.deposit(100)
```

client q

```
ta = a.getBalance()
↓
tb = b.getBalance()
↓
Total = t1 + tb
```

10 / 42

Serial equivalence

The isolation requirement states that the outcome of a set of transactions should be the same as the outcome when the transactions are executed in sequence.

We call this *serial equivalence*.

Should we abandon all hope of executing transactions concurrently?

11 / 42

conflicting operations

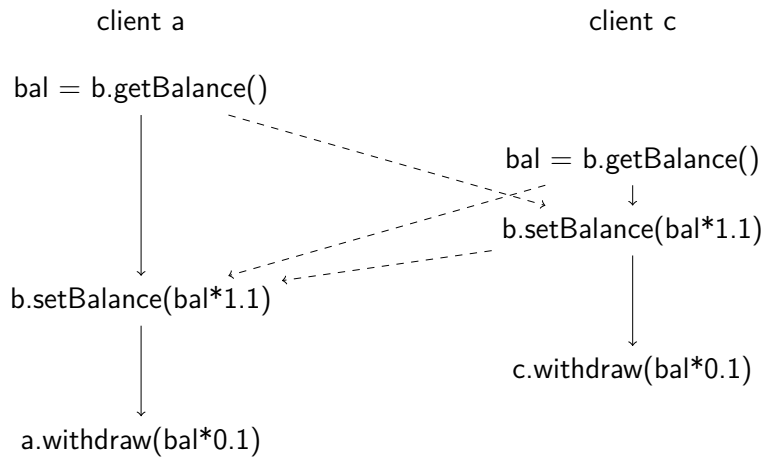
Which operations are order sensitive?

- read - read
- read - write
- write - write

Two transactions are serially equivalent if, and only if, all pair of conflicting operations of the transactions are executed *in the same order*.

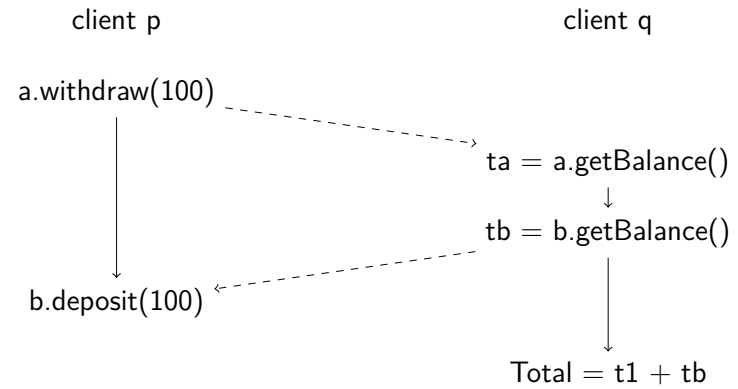
12 / 42

Lost update - revisited



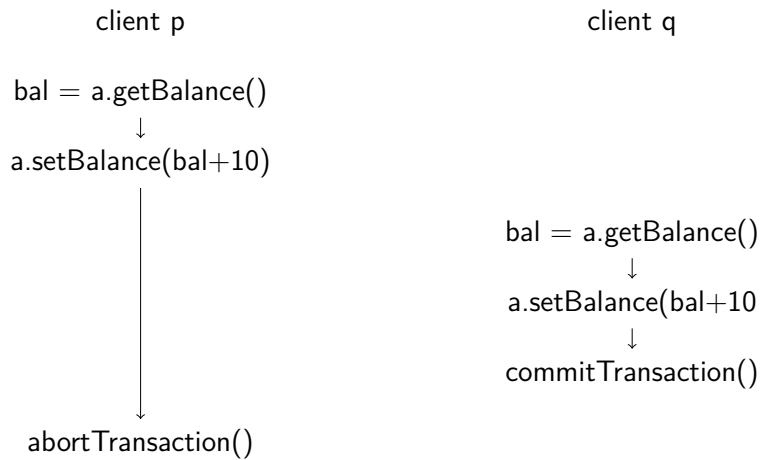
13 / 42

inconsistent retrieval - revisited



14 / 42

more problems



15 / 42

Recoverability

In order to recover from an aborting transaction: a transaction must not commit if it has done a *dirty read*.

16 / 42

Assume we do a *dirty read*, write values and then wait to commit.

A second process, reads our dirty values, writes values and waits to commit.

A third process, reads the dirty values, writes values and waits to commit.

...

We abort

In order to avoid cascading aborts we should suspend when we read a dirty value.

- To be recoverable a transaction must suspend its commit operation if it has performed a dirty read.
- If a transaction aborts, any suspended transaction must be aborted.
- To prevent cascading aborts, a transaction could be prevented from performing a read operation of a non-committed value.
 - Once the value is committed or the previous transaction aborts the execution can continue.
 - We will restrict concurrency.

17 / 42

18 / 42

client p

a.setBalance(105)



abortTransaction()

client q

a.setBalance(110)



commitTransaction()

Also write operations must be delayed in order to be able to recover from an aborting transaction.

- In general, both read and write operations must be delayed until all previous transactions containing write operations have been aborted or committed.
- Strict execution enforces isolation, no visible effects until commit.
- How do we implement strict execution efficiently?

19 / 42

20 / 42

...increase concurrency while preserving serial equivalence?

- locking: simple but dangerous
- optimistic: large overhead if many conflicts
- timestamp: ok, if time would be simple

Idea - lock all objects to prevent other transaction to read from or write to the same objects.

To guarantee serial equivalence we require two phase locking:

- lock objects in any order,
- release locks in any order,
- commit

We are not allowed to take a lock if a lock has been released.

Does not handle the problem with dirty read and premature write.

- To handle dirty read and premature write:
 - lock in any order
 - commit or abort
 - unlock
- Can we increase concurrency?

- two-version locking: read and write
- allow multiple readers but only one writer
- promote read locks to write locks
- strict two-phase locking prevents demotion

Two-version locking

Similar idea but now with: read, write and commit locks.

- A read lock is allowed unless a *commit lock* is taken.
- One write lock is allowed if no commit lock is taken (i.e. even if read locks are taken)
- Written values are held local to the transaction and are not visible before commit.
- A write lock can be promoted to a commit lock if there are no read locks.
- When a transaction commits it tries to promote write locks to commit locks.

25 / 42

Hierarchical locks

Idea: locks of mixed granularity.

- Small locks increase concurrency
- Large locks decrease overhead

26 / 42

Why locking s*cks

- Locking is an overhead not present in a non-concurrent system. You're paying even if there is no conflict.
- There is always the risk of deadlock or the locking scheme is so restricted that it prevents concurrency.
- To avoid cascading aborts, locks must be held to the end of the transaction.

27 / 42

Optimistic concurrency control

- Perform transaction in a copy of an object, hoping that no other transaction will interfere.
- When performing a commit operation *the validity* is controlled.
- If transaction is *valid*, the values written to permanent storage.

28 / 42

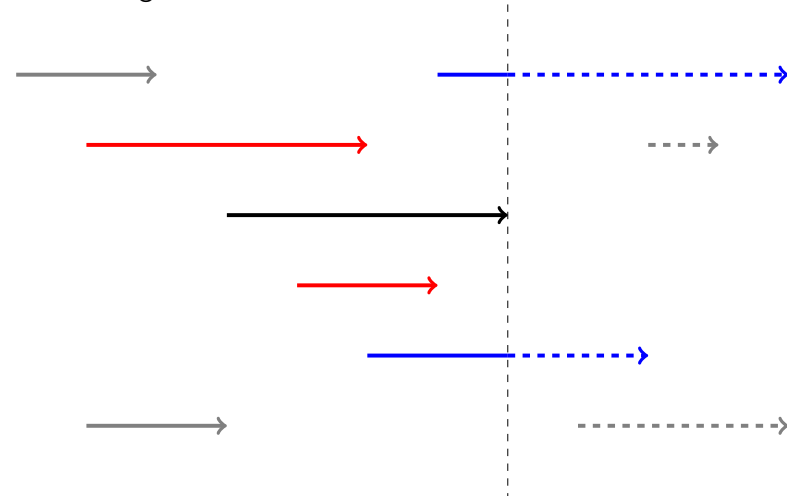
Let's be optimistic

- If we are lucky, transactions do not have any conflicting operations.
- The validity check is quick and successful.
- The update phase is simple.

29 / 42

Validation

Like driving a car in Damascus.

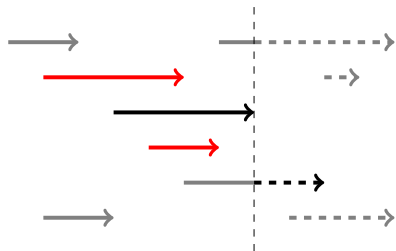


30 / 42

Backwards validation

Validate a transaction by comparing all:

- read operations with committed write operations
- if a conflict is found, abort

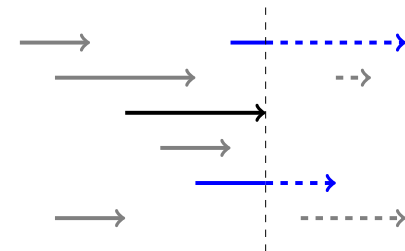


31 / 42

Forward validation

Validate a transaction by comparing all:

- write operations with conflicting read operations
- if a conflict is found, abort
- ...
- ... or, kill the other transaction



32 / 42

Works well if there are no conflicts.

- Backward validation: simpler to implement, need to save all write operations
- Forward validation: moving target, flexible if not successful

How do we guarantee liveness?

Each transaction is given a time stamp when started.

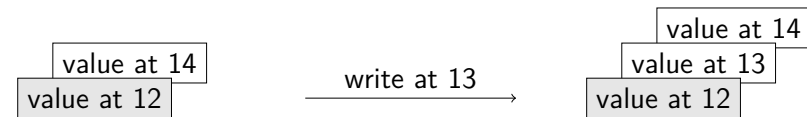
Operations are validated when performed:

- writing only if no later transaction has read or written
- reading only if no later transaction has written

Hmm, requires some bookkeeping.

Each objects keep a list of *tentative*, not committed, versions of the value.

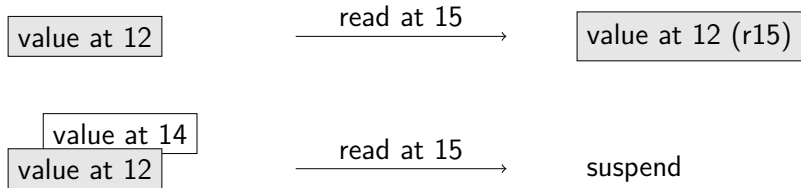
- Write operations can be inserted in the right order, no fear for deadlocks.
- Read operations wait for tentative values to be committed.
- If an operation **arrives too late** the transaction is aborted.



A sequence of tentative values, not yet committed.

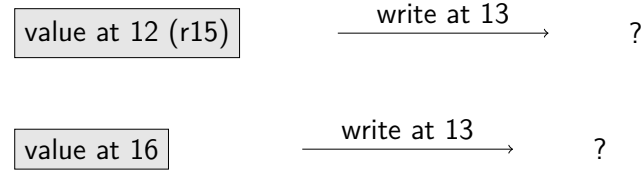


tentative values - read



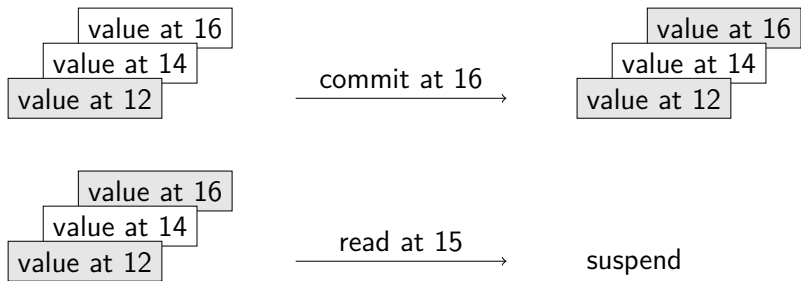
37 / 42

tentative values - what to do?



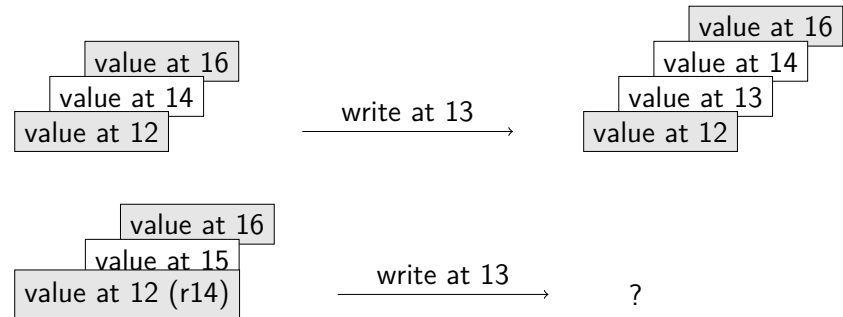
38 / 42

tentative values - how about this



39 / 42

multiversion timestamp



40 / 42

- consistency is checked when the operation is performed
- commit is always successful
- an operation can suspend or arrive too late
- read operations will succeed, suspend or arrive too late
- write operations will succeed or arrive too late
- multiversion timestamp can improve performance

Transactions group sequences of operations into a ACID operation.

- Atomic: all or nothing
 - Consistent: leave the server in a consistent state
 - Isolation: same result as having executed in sequence
 - Durability: safe even if server crashes
 - problem is how to increase concurrency
 - need to preserve serial equivalence
 - aborting transactions is a problem
 - how do we maximize concurrency
- Implementations: locking, optimistic concurrency control, timestamps