Many of the slides through slide 21 are figures from Coulouris, Dollimore, Kindberg, and Blair.
Transactions

• “ACID” property:
  – atomic
    - all or nothing
  – consistent
    - take system from one consistent state to another
  – isolated
    - have no effect on other transactions until committed
  – durable
    - persists
Figure 16.1. A pair of interfaces to be used in upcoming examples.

**Operations of the Account interface**

- deposit(amount)
  deposit amount in the account
- withdraw(amount)
  withdraw amount from the account
- getBalance() \rightarrow \text{amount}
  return the balance of the account
- setBalance(amount)
  set the balance of the account to amount

---

**Operations of the Branch interface**

- create(name) \rightarrow \text{account}
  create a new account with a given name
- lookUp(name) \rightarrow \text{account}
  return a reference to the account with the given name
- branchTotal() \rightarrow \text{amount}
  return the total of all the balances at the branch
A client’s banking transaction

Transaction $T$:
- $a$.withdraw(100);
- $b$.deposit(100);
- $c$.withdraw(200);
- $b$.deposit(200);

Figure 16.2: A sample banking transaction using the account interface.
Operations in *Coordinator* interface

`openTransaction() -> trans;`
starts a new transaction and delivers a unique TID

`trans. This identifier will be used in the other
operations in the transaction.

`closeTransaction(trans) -> (commit, abort);`
ends a transaction: a *commit* return value indicates
that the transaction has committed; an *abort* return
value indicates that it has aborted.

`abortTransaction(trans);`
aborts the transaction.

Figure 16.3
### Transaction Life Histories

<table>
<thead>
<tr>
<th>Successful</th>
<th>Aborted by client</th>
<th>Aborted by server</th>
</tr>
</thead>
<tbody>
<tr>
<td>openTransaction</td>
<td>openTransaction</td>
<td>openTransaction</td>
</tr>
<tr>
<td>operation</td>
<td>operation</td>
<td>operation</td>
</tr>
<tr>
<td>operation</td>
<td>operation</td>
<td>operation</td>
</tr>
<tr>
<td>closeTransaction</td>
<td>abortTransaction</td>
<td></td>
</tr>
</tbody>
</table>

- server aborts transaction
- operation ERROR reported to client
The lost update problem

Initial balances: a: $100, b: $200, c: $300

<table>
<thead>
<tr>
<th>Transaction T:</th>
<th>Transaction U:</th>
</tr>
</thead>
<tbody>
<tr>
<td>balance = b.getBalance();</td>
<td>balance = b.getBalance();</td>
</tr>
<tr>
<td>b.setBalance(balance*1.1);</td>
<td>b.setBalance(balance*1.1);</td>
</tr>
<tr>
<td>a.withdraw(balance/10)</td>
<td>c.withdraw(balance/10)</td>
</tr>
</tbody>
</table>

| balance = b.getBalance(); | $200 |
| b.setBalance(balance*1.1); | $220 |
| a.withdraw(balance/10) | $80 |
| c.withdraw(balance/10) | $280 |

Figure 16.5
The inconsistent retrievals problem

<table>
<thead>
<tr>
<th>Transaction V:</th>
<th>Transaction W:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a.\text{withdraw}(100))</td>
<td>(a\text{Branch}.\text{branchTotal}())</td>
</tr>
<tr>
<td>(b.\text{deposit}(100))</td>
<td></td>
</tr>
<tr>
<td>(a.\text{withdraw}(100)); $100</td>
<td>(\text{total} = a.\text{getBalance}()); $100</td>
</tr>
<tr>
<td></td>
<td>(\text{total} = \text{total}+b.\text{getBalance}()); $300</td>
</tr>
<tr>
<td>(b.\text{deposit}(100)); $300</td>
<td>(\text{total} = \text{total}+c.\text{getBalance}()); $0</td>
</tr>
</tbody>
</table>

Figure 16.6
Serial Equivalence

- Consider the effect of a concurrent execution of transactions A and B
  \[ A \parallel B \]
- What should our correctness criteria be?
- Intuitively, it should be equivalent to some serial execution:
  \[ A;B \]
  \[ B;A \]
- We say that \( A \parallel B \) is *serially equivalent* if it has the same effect as either \( A;B \) or \( B;A \)
### A serially equivalent interleaving of $T$ and $U$

<table>
<thead>
<tr>
<th>Transaction $T$:</th>
<th>Transaction $U$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$balance = b.getBalance()$</td>
<td>$balance = b.getBalance()$</td>
</tr>
<tr>
<td>$b.setBalance(balance*1.1)$</td>
<td>$b.setBalance(balance*1.1)$</td>
</tr>
<tr>
<td>$a.withdraw(balance/10)$</td>
<td>$c.withdraw(balance/10)$</td>
</tr>
<tr>
<td>$balance = b.getBalance()$ $200$</td>
<td>$balance = b.getBalance()$ $220$</td>
</tr>
<tr>
<td>$b.setBalance(balance*1.1)$ $220$</td>
<td>$b.setBalance(balance*1.1)$ $242$</td>
</tr>
<tr>
<td>$a.withdraw(balance/10)$ $80$</td>
<td>$c.withdraw(balance/10)$ $278$</td>
</tr>
</tbody>
</table>

---

Figure 16.7
### A serially equivalent interleaving of V and W

<table>
<thead>
<tr>
<th>Transaction V:</th>
<th>Transaction W:</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{a.withdraw(100);}</td>
<td></td>
</tr>
<tr>
<td>\texttt{b.deposit(100)}</td>
<td>\texttt{aBranch.branchTotal()}</td>
</tr>
<tr>
<td>\texttt{a.withdraw(100)};</td>
<td>$100</td>
</tr>
<tr>
<td>\texttt{b.deposit(100)}</td>
<td>$300</td>
</tr>
</tbody>
</table>

total = \texttt{a.getBalance()} $100  
\texttt{total} = \texttt{total} + b.\texttt{getBalance()} $400  
\texttt{total} = \texttt{total} + c.\texttt{getBalance()}  
\ldots

---

Figure 16.8
Read and write operation conflict rules

<table>
<thead>
<tr>
<th>Operations of different transactions</th>
<th>Conflict</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>read read</td>
<td>No</td>
<td>Because the effect of a pair of read operations does not depend on the order in which they are executed</td>
</tr>
<tr>
<td>read write</td>
<td>Yes</td>
<td>Because the effect of a read and a write operation depends on the order of their execution</td>
</tr>
<tr>
<td>write write</td>
<td>Yes</td>
<td>Because the effect of a pair of write operations depends on the order of their execution</td>
</tr>
</tbody>
</table>

Figure 16.9
Figure 16.10: Note that all of R’s accesses to i come before S’s, and all of S’s accesses to j come before R’s. Nevertheless this is not a serially equivalent concurrent execution of R and S.
Concurrency Control

- How to ensure that transaction executions are serializable:
  - locking
  - optimism
Figure 16.11: The isolation property requires that transactions do not see the uncommitted state of other transactions.
Figure 16.12: Suppose U aborts after overwriting A.
Transactions $T$ and $U$ with exclusive locks

<table>
<thead>
<tr>
<th>Transaction $T$:</th>
<th>Transaction $U$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$balance = b.getBalance()$</td>
<td>$balance = b.getBalance()$</td>
</tr>
<tr>
<td>$b.setBalance(bal*1.1)$</td>
<td>$b.setBalance(bal*1.1)$</td>
</tr>
<tr>
<td>$a.withdraw(bal/10)$</td>
<td>$c.withdraw(bal/10)$</td>
</tr>
<tr>
<td>Operations</td>
<td>Operations</td>
</tr>
<tr>
<td>Locks</td>
<td>Locks</td>
</tr>
</tbody>
</table>

- openTransaction
  - $bal = b.getBalance()$: lock B
- $b.setBalance(bal*1.1)$: wait for $T$'s lock on B
- $a.withdraw(bal/10)$: lock A
- closeTransaction: unlock A, B

- openTransaction: wait for $T$'s lock on B
- $bal = b.getBalance()$: wait for $T$'s lock on B
- $b.setBalance(bal*1.1)$: lock B
- $c.withdraw(bal/10)$: lock C
- closeTransaction: unlock B, C

---

Figure 16.14
Two-Phase Locking

1) Acquire locks
2) Release locks

- No more locks may be acquired after any lock is released
- Strict two-phase locking
  - no locks released until transaction commits
Transaction A completes before transaction B. Two-phase locking was used. Is it necessarily the case that their concurrent execution is equivalent to first executing all of A, then all of B?
Transaction Steps

- Accumulate changes
  - store as “tentative versions”
- Make sure everything is ok
- Commit or abort
  - move tentative versions to actual
  or
  - delete tentative versions
Figure 16.13
Much of the remainder of this lecture is adapted from the textbook by Tanenbaum and Van Steen and from Chapter 7 of *Concurrency Control and Recovery in Database Systems*, by P. Bernstein, V. Hadzilacos, and N. Goodman, Addison-Wesley (1987). The latter text is available at http://research.microsoft.com/en-us/people/philbe/ccontrol.aspx.
Cooperation

Coordinator

Begin Transaction:
- a.withdraw(100);
- b.deposit(50);
- c.deposit(50);
End Transaction;

client

withdraw(100);

a

deposit(50);

b

deposit(50);

c
Atomic Commit

- AC1: All participants that reach a (commit/abort) decision reach the same one
- AC2: A participant cannot reverse its decision
- AC3: The commit decision can be reached only if all participants agree
- AC4: If there are no failures and all participants vote yes, then decision will be commit
- AC5: For any execution, if all failures are repaired and no new failures occur for a sufficiently long interval, then all participants will reach a (commit/abort) decision

This is adapted from Bernstein et al.
Two-Phase Commit

• Phase 1
  – coordinator prepares to commit:
    - asks participants to vote either “commit” or “abort”
  – participants respond appropriately

• Phase 2
  – coordinator decides outcome:
    - if all participants vote commit, outcome is commit, otherwise outcome is abort
    - outcome sent to all participants
  – participants do what they’re told
The labeling of the arcs (A/B) means that if “A” occurs, then perform action “B” and follow the arc to the next state.
Failures

- Coordinator or participants could crash
  - assume “fail-stop”
    - crash detected by time-out
    - no byzantine failures
  - crashed machines restart
    - recover their state
Crash Points

Coordinator

- Init
  - app commit/vote req
  - any abort/abort
  - all commit/commit

Wait

- Abort
- Commit

Participant

Init

- vote req/abort

Uncertain

- abort/ack
- commit/ack

- Abort
- Commit
Dealing with Timeouts (1)

- Coordinator times out in *Wait* state
  - waiting for a participant to vote
  - takes no response to mean "abort"
  - sends abort to all other participants
- Participant times out in *Uncertain* state
  - waiting for coordinator to say "commit" or "abort"
  - can’t assume either outcome
  - waits for coordinator to restart
  - contacts coordinator for final outcome
This implies that the participants know one another’s identities. They could be supplied by the coordinator in the initial vote request.
Improving on Two-Phase Commit

- It works fine in practice!
- But …
  - all participants could conceivably be in uncertain state and coordinator is down (for a long time)
- Can we make it so such blocking can’t happen?
What Causes Blocking?

- Coordinator is down
- If all operational (not-failed) participants are in uncertain state, they are blocked
- If all participants are operational, they can elect new coordinator
- If any participant has crashed, the others don’t know if it crashed before or after voting (to commit or abort)
Guaranteeing Non-Blocking

- Non-blocking property (NB):
  - *if any operational process is in the Uncertain state, then no process (operational or failed) can have decided to commit*

- If NB holds, then operational processes may elect new coordinator and decide to commit or abort

Note that NB does not hold for two-phase commit!
For our upcoming discussion of three-phase commit, we assume that the only sort of failure is that of a machine crashing (then recovering). In particular, communication failures do not happen.
Three-Phase Commit

• Phase 1
  – coordinator prepares to commit:
    - asks participants to vote either “commit” or “abort”
    – participants respond appropriately
• Phase 2
  – coordinator counts votes:
    - if all participants vote commit, outcome is pre-commit, otherwise outcome is abort
    - outcome sent to all participants
    – participants ack and either abort or wait for commit
• Phase 3
  – coordinator waits for all acks
    - if committing, sends final commit to all participants
    – participants commit
Revised State Diagrams

Init
  ↓
  app commit/vote req

Wait
  ↓
  any abort/abort

Abort
  ↓
  all commit/precommit

Pre Commit
  ↓
  all ack/commit

Commit
  ↓

Init
  ↓
  vote req/abort

Uncertain
  ↓
  abort/ack

Abort
  ↓
  precommit/ack

Pre Commit
  ↓
  commit/commit

Commit
If a participant times out in its init state while waiting for a vote req from the coordinator, it may safely unilaterally abort.
If the coordinator times out in its wait state while waiting to receive votes from participants, it should send aborts to all operational participants.
If the coordinator times out while in its precommit state, waiting to receive acks from the participants, it may safely commit, since it had received commit votes from all. The failed participants will learn about the commit when they reboot.
If a participant times out in its uncertain state waiting to hear from the coordinator, it must communicate with the other operational participants to determine if it should commit or abort. In particular, if any other participant has aborted, it should abort. But what if this is not the case? (Go on to the next slides ...)
If a participant times out in its precommit state, waiting to hear from the coordinator, shouldn’t it simply assume it may commit? The answer is no, because that might violate NB: there may be some other participant that’s still in the uncertain state.

The situation we’re concerned about is that, after committing, the participant might fail, while some other participant (perhaps the only other participant) remains operational, but in the uncertain state. That participant, now not knowing anything about the states of the others, should be allowed to abort by virtue of NB.
Details (1)

• If original coordinator remains operational
  – participant crashes handled as in two-phase commit

• If participant times out in *Uncertain* or *PreCommit* states
  – if any other participant has aborted, it aborts (it must have been in Uncertain state)
  – otherwise, it starts an election for a new coordinator
Details (2)

- When newly elected coordinator starts up
  - sends state-request message to all operational participants
  - coordinator collects states and proceeds according to four termination rules (termination protocol):
    - TR1: if any participant is in Abort state, all are sent abort messages
    - TR2: if some participant is in Commit state, all are sent commit messages
    - TR3: if all participants are in Uncertain state, all are sent abort messages
    - TR4: if some participant is in PreCommit state, but none in Commit State, those in Uncertain state are sent PreCommit messages; once these are acked, all participants are sent commit messages

Note that the newly elected coordinator could fail. If so, a new one is elected. (Participants will time-out waiting for a message.)
Details (3)

- When failed participant comes back up
  - if it failed in Init state
    - it aborts
  - Otherwise it asks other participants for outcome
    - will eventually get either commit or abort
      - (could get abort even if it was in the PreCommit state when it crashed)
Correctness (1)

- Lemma 1: After a new coordinator starts up, exactly one of TR1 – TR4 will hold
- Theorem 1: In the absence of total failures, participants will never block
  - they clearly won’t block if the coordinator never fails
  - if the coordinator fails, a new one is elected
  - one of TR1-TR4 will hold and a decision will be reached
  - if the new coordinator fails, a new one is elected; if it fails another is elected, etc. until there are no more participants

For details, see chapter 7 from Bernstein et al.
Correctness (2)

- Lemma 2: All participants that reach a decision on the same invocation of the termination protocol reach the same one
- Lemma 3: If NB holds before the termination protocol starts, it holds through the execution of the protocol
- Theorem 2: All operational participants reach the same decision
  - proof by induction on the invocations of the termination protocol
Total Failure

- What if coordinator and all participants fail?
- When they come back up, how do they decide?
  - if resurrected participant either didn’t vote or voted abort, it may unilaterally abort
  - otherwise, must run termination protocol
  - but works only if last participant to fail has come back up

See Bernstein et al. for details.
Communication Failures

- Network could partition into multiple pieces
- Not sufficient to get agreement in a piece containing a quorum
  - consensus is required for commit!
- Scenario
  - all participants vote
  - coordinator collects results
  - network partitions before or after all results collected
  - if network reconnects: easy
  - network never fully reconnects, but each participant eventually can communicate (perhaps briefly) with all others

See Bernstein et al. for details.