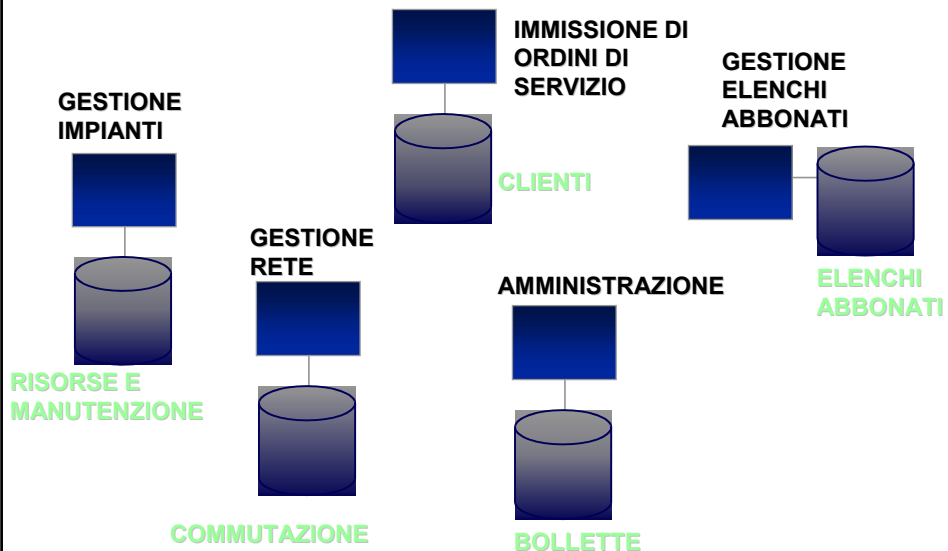
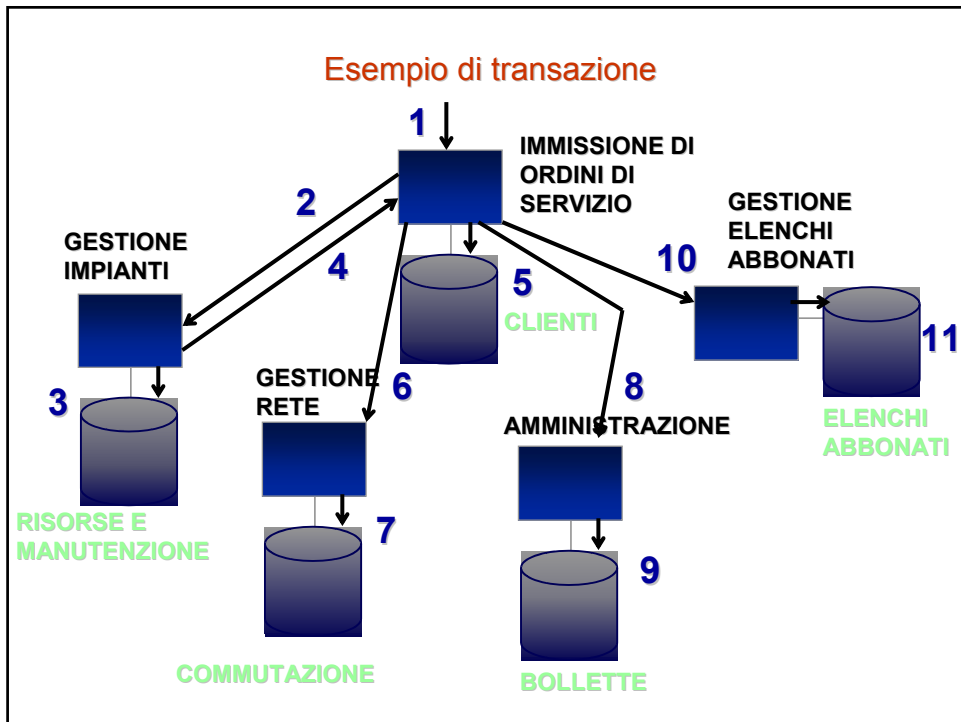


**Basi di dati vol.2**  
**Capitolo 2**  
**Gestione delle transazioni**

12/10/2003

**Esempio di sistema informativo**

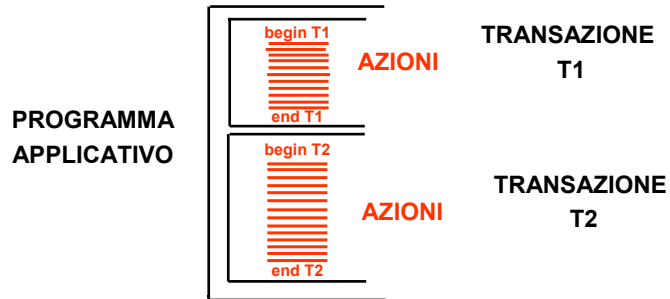




## DEFINIZIONE DI TRANSAZIONE

- Transazione: parte di programma caratterizzata da un inizio (**begin-transaction**, *start transaction* in SQL), una fine (**end-transaction**, *non esplicitata in SQL*) e al cui interno deve essere eseguito una e una sola volta uno dei seguenti comandi
  - **commit work** per terminare correttamente
  - **rollback work** per abortire la transazione
- Un **sistema transazionale (OLTP)** e' in grado di definire ed eseguire transazioni per conto di un certo numero di applicazioni concorrenti

## DIFFERENZA FRA APPLICAZIONE E TRANSAZIONE



## Una transazione

```
start transaction;  
update ContoCorrente  
  set Saldo = Saldo + 10 where NumConto = 12202;  
update ContoCorrente  
  set Saldo = Saldo - 10 where NumConto = 42177;  
commit work;
```

## Una transazione con varie decisioni

```
start transaction;
update ContoCorrente
  set Saldo = Saldo + 10 where NumConto = 12202;
update ContoCorrente
  set Saldo = Saldo - 10 where NumConto = 42177;
select Saldo into A
  from ContoCorrente
  where NumConto = 42177;
if (A>=0) then commit work
  else rollback work;
```

## Transazioni in JDBC

- Scelta della modalità delle transazioni: un metodo definito nell'interfaccia `Connection`:  
`setAutoCommit(boolean autoCommit)`
- `con.setAutoCommit(true)`
  - (default) "autocommit": ogni operazione è una transazione
- `con.setAutoCommit(false)`
  - gestione delle transazioni da programma  
`con.commit()`  
`con.rollback()`
  - non c'è `start transaction`

## Il concetto di transazione

- Una unità di elaborazione che gode delle proprietà "ACIDE"
  - Atomicità
  - Consistenza
  - Isolamento
  - Durata (persistenza)

## Atomicità

- Una transazione è una unità atomica di elaborazione
- Non può lasciare la base di dati in uno stato intermedio
  - un guasto o un errore prima del commit debbono causare l'annullamento (UNDO) delle operazioni svolte
  - un guasto o errore dopo il commit non deve avere conseguenze; se necessario vanno ripetute (REDO) le operazioni
- Esito
  - Commit = caso "normale" e più frequente (99% ?)
  - Abort (o rollback)
    - richiesto dall'applicazione = suicidio
    - requested dal sistema (violazione dei vincoli, concorrenza, incertezza in caso di fallimento) = omicidio

## Consistenza

- La transazione rispetta i vincoli di integrita'
- Conseguenza:
  - se lo stato iniziale e' corretto
  - anche lo stato finale e' corretto

## Isolamento

- La transazione non risente degli effetti delle altre transazioni concorrenti
  - l'esecuzione concorrente di una collezione di transazioni deve produrre un risultato che si potrbbe ottenerre con una esecuzione sequenziale
- Conseguenza: una transazione non espone i suoi stati intermedi
  - Si evita l' "effetto domino"

## Durabilità (Persistenza)

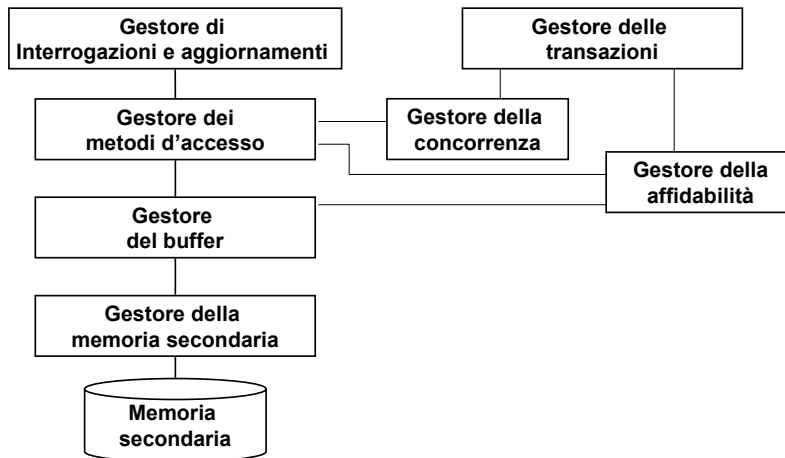
- Gli effetti di una transazione andata in commit non vanno perduti ("durano per sempre"), anche in presenza di guasti
  - Commit significa impegno

## Transazioni e moduli di DBMS

- Atomicità e durabilità
  - Gestore dell'affidabilità (Reliability manager)
- Isolamento:
  - Gestore della concorrenza
- Consistenza:
  - Gestore dell'integrità a tempo di esecuzione (con il supporto del compilatore del DDL)

## Gestore degli accessi e delle interrogazioni

## Gestore delle transazioni

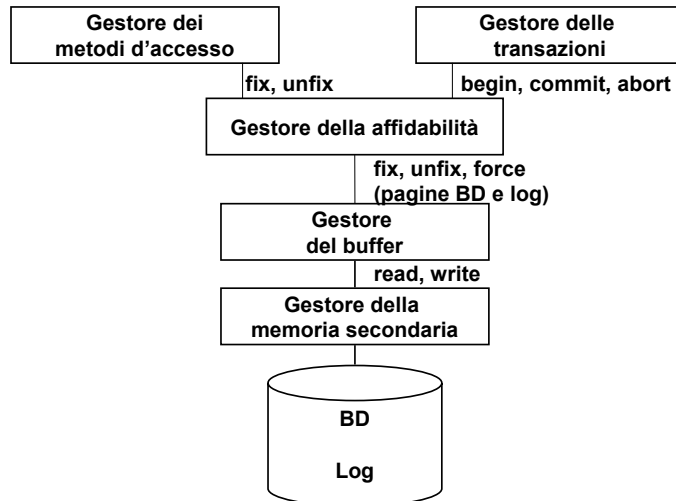


## Gestore dell'affidabilità

- Gestisce l'esecuzione dei comandi transazionali
  - `start transaction (B, begin)`
  - `commit work (C)`
  - `rollback work (A, abort)`
- e le operazioni di ripristino (recovery) dopo i guasti :
  - *warm restart* e *cold restart*
- Assicura atomicità e durabilità
- Usa il **log**:
  - Un archivio permanente che registra le operazioni svolte
  - Due metafore: il filo di Arianna e i sassolini e le briciole di Hansel e Gretel



## Architettura del controllore dell'affidabilità



12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

17

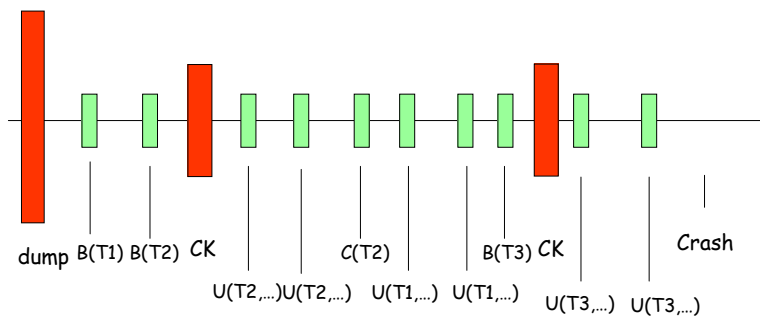
## Persistenza delle memorie

- **Memoria centrale**: non e' persistente
- **Memoria di massa**: e' persistente ma puo' danneggiarsi
- **Memoria stabile**: memoria che non puo' danneggiarsi (e' una astrazione):
  - perseguita attraverso la ridondanza:
    - dischi replicati
    - nastri
    - ...

## Il log

- Il log è un file sequenziale gestito dal controllore dell'affidabilità, scritto in memoria stabile
- "Diario di bordo": riporta tutte le operazioni in ordine
- Record nel log
  - *operazioni delle transazioni*
    - begin, B(T)
    - insert, I(T,O,AS)
    - delete, D(T,O,BS)
    - update, U(T,O,BS,AS)
    - commit, C(T), abort, A(T)
  - *record di sistema*
    - dump
    - checkpoint

## Struttura del log



## Log, checkpoint e dump: a che cosa servono?

- Il log serve "a ricostruire" le operazioni
- Checkpoint e dump servono ad evitare che la ricostruzione debba partire dall'inizio dei tempi
  - si usano con riferimento a tipi di guasti diversi

## Undo e redo

- Undo di una azione su un oggetto  $O$ :
  - update, delete: copiare il valore del **before state** ( $BS$ ) nell'oggetto  $O$
  - insert: eliminare  $O$
- Redo di una azione su un oggetto  $O$ :
  - insert, update: copiare il valore dell' **after state** ( $AS$ ) into nell'oggetto  $O$
  - delete: reinserire  $O$
- *Idempotenza* di *undo* e *redo*:
  - $undo(undo(A)) = undo(A)$
  - $redo(redo(A)) = redo(A)$

## Checkpoint

- Operazione che serve a "fare il punto" della situazione, semplificando le successive operazioni di ripristino:
  - ha lo scopo di registrare quali transazioni sono attive in un certo istante (e dualmente, di confermare che le altre o non sono iniziate o sono finite)
- Paagone (estremo):
  - la "chiusura dei conti" di fine anno di una amministrazione:
    - dal 25 novembre (ad esempio) non si accettano nuove richieste di "operazioni" e si concludono tutte quelle avviate prima di accettarne di nuove

## Checkpoint (2)

- Varie modalità, vediamo la più semplice:
  - si sospende l'accettazione di richieste di ogni tipo (scrittura, inserimenti, ..., commit, abort)
  - si trasferiscono in memoria di massa (tramite *force*) tutte le pagine sporche relative a transazioni andate in commit
  - si registrano sul log in modo sincrono (*force*) gli identificatori delle transazioni in corso
  - si riprende l'accettazione delle operazioni
- Così siamo sicuri che
  - per tutte le transazioni che hanno effettuato il commit i dati sono in memoria di massa
  - le transazioni "a metà strada" sono elencate nel checkpoint

## Dump

- Copia completa ("di riserva", backup) della base di dati
  - Solitamente prodotta mentre il sistema non è operativo
  - Salvato in memoria stabile, come
  - The copy is stored in the stable memory, typically on tape, and is called *backup*
  - Un record di `dump` nel log indica il momento in cui il log è stato effettuato (e dettagli pratici, file, dispositivo, ...)

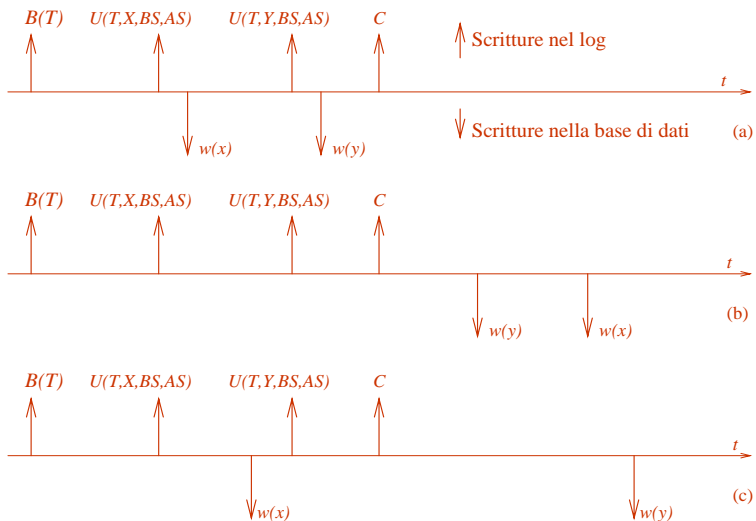
## Esito di una transazione

- L'esito di una transazione è determinato irrevocabilmente quando viene scritto il record di **commit** nel log in modo sincrono, con una **force**
  - un guasto prima di tale istante porta ad un undo di tutte le azioni, per ricostruire lo stato originario della base di dati
  - un guasto successivo non deve avere conseguenze: lo stato finale della base di dati deve essere ricostruito, con redo se necessario
- record di `abort` possono essere scritti in modo asincrono

## Regole fondamentali per il log

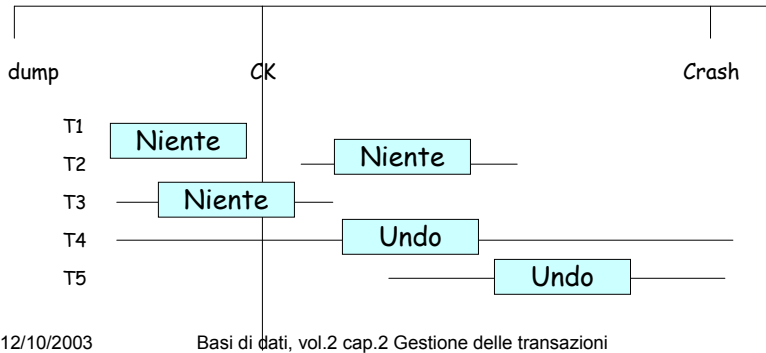
- **Write-Ahead-Log:**
  - si scrive il giornale (parte before) prima del database
    - consente di disfare le azioni
- **Commit-Precedenza:**
  - si scrive il giornale (parte after) prima del commit
    - consente di rifare le azioni
- Quando scriviamo nella base di dati?
  - Varie alternative

## Scrittura nel log e nella base di dati



## Modalita' immediata

- Il DB contiene valori AS provenienti da transazioni uncommitted
- Richiede Undo delle operazioni di transazioni uncommitted al momento del guasto
- Non richiede Redo



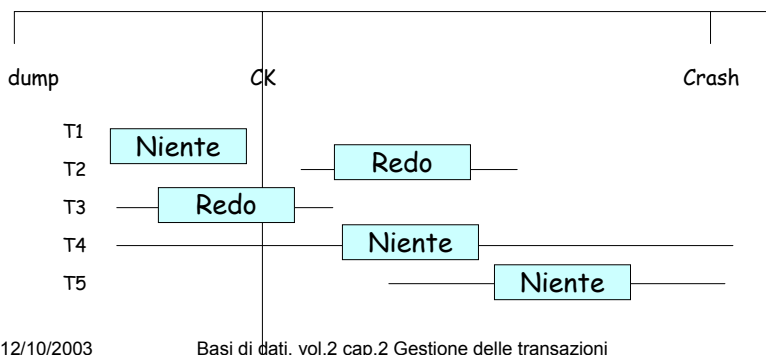
12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

29

## Modalita' differita

- Il DB non contiene valori AS provenienti da transazioni uncommitted
- In caso di abort, non occorre fare niente
- Rende superflua la procedura di Undo. Richiede Redo



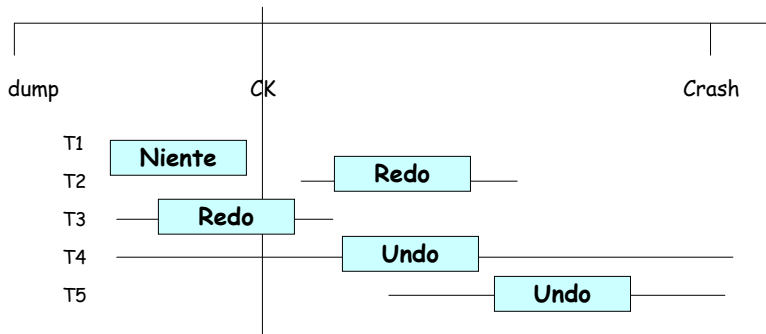
12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

30

## Essite una terza modalita': modalità mista

- La scrittura puo' avvenire in modalita' sia immediata che differita
- Consente l'ottimizzazione delle operazioni di flush
- Richiede sia Undo che Redo



12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

31

## Optimizations

- Several log records are normally written into the same log page
- Several log records of the same transactions can be forced at commit time
- Several transactions can commit together by forcing their `commit` record with the same operation (*group commit*)
- Systems can exploit parallelism in writing logs

12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

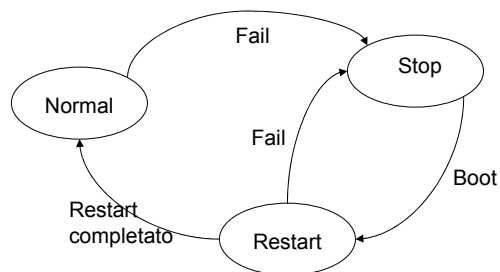
32



## Guasti

- **Guasti "soft"**: errori di programma, crash di sistema, caduta di tensione
  - si perde la memoria centrale
  - non si perde la memoria secondaria**warm restart, ripresa a caldo**
- **Guasti "hard"**: sui dispositivi di memoria secondaria
  - si perde anche la memoria secondaria
  - non si perde la memoria stabile (e quindi il log)**cold restart, ripresa a freddo**

## Modello "fail-stop"



## Processo di restart

- Obiettivo: classificare le transazioni in
  - completate (tutti i dati in memoria stabile)
  - in commit ma non necessariamente completate (può servire redo)
  - senza commit (vanno annullate, undo)

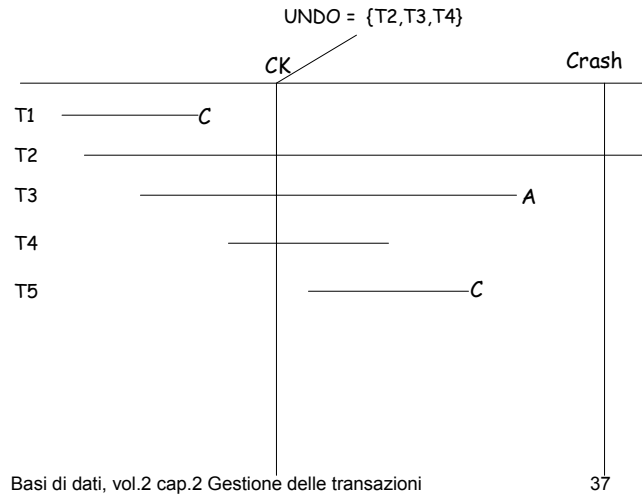
## Ripresa a caldo

Quattro fasi:

- trovare l'ultimo checkpoint (ripercorrendo il log a ritroso)
- costruire gli insiemi *UNDO* (transazioni da disfare) e *REDO* (transazioni da rifare)
- ripercorrere il log all'indietro, fino alla più vecchia azione delle transazioni in *UNDO* e *REDO*, disfacendo tutte le azioni delle transazioni in *UNDO*
- ripercorrere il log in avanti, rifacendo tutte le azioni delle transazioni in *REDO*

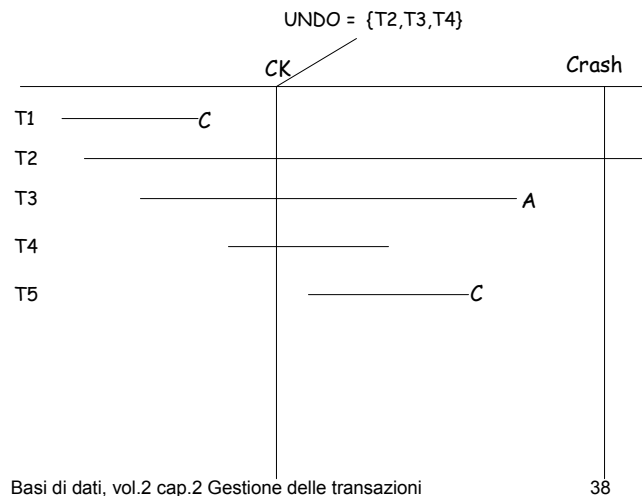
## Esempio di warm restart

B(T1)  
 B(T2)  
 U(T2, O1, B1, A1)  
 I(T1, O2, A2)  
 B(T3)  
 C(T1)  
 B(T4)  
 U(T3, O2, B3, A3)  
 U(T4, O3, B4, A4)  
 CK(T2, T3, T4)  
 C(T4)  
 B(T5)  
 U(T3, O3, B5, A5)  
 U(T5, O4, B6, A6)  
 D(T3, O5, B7)  
 A(T3)  
 C(T5)  
 I(+2, O6, A8)  
 12/10/2003



## 1. Ricerca dell'ultimo checkpoint

B(T1)  
 B(T2)  
 U(T2, O1, B1, A1)  
 I(T1, O2, A2)  
 B(T3)  
 C(T1)  
 B(T4)  
 U(T3, O2, B3, A3)  
 U(T4, O3, B4, A4)  
 CK(T2, T3, T4)  
 C(T4)  
 B(T5)  
 U(T3, O3, B5, A5)  
 U(T5, O4, B6, A6)  
 D(T3, O5, B7)  
 A(T3)  
 C(T5)  
 I(+2, O6, A8)  
 12/10/2003



## 2. Costruzione degli insiemi UNDO e REDO

B(T1)	0. UNDO = {T2,T3,T4}. REDO = {}
B(T2)	<hr/>
8. U(T2, O1, B1, A1)	1. C(T4) → UNDO = {T2, T3}. REDO = {T4}
I(T1, O2, A2)	2. B(T5) → UNDO = {T2,T3,T5}. REDO = {T4} Setup
B(T3)	3. C(T5) → UNDO = {T2,T3}. REDO = {T4, T5}
C(T1)	
B(T4)	
7. U(T3,O2,B3,A3)	
9. U(T4,O3,B4,A4)	
CK(T2,T3,T4)	
1. C(T4)	
2. B(T5)	
6. U(T3,O3,B5,A5)	
10. U(T5,O4,B6,A6)	
5. D(T3,O5,B7)	
A(T3)	
3. C(T5)	
4. I(T2,O6,A8)	

12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

39

## 3. Fase UNDO

B(T1)	0. UNDO = {T2,T3,T4}. REDO = {}
B(T2)	<hr/>
8. U(T2, O1, B1, A1)	1. C(T4) → UNDO = {T2, T3}. REDO = {T4}
I(T1, O2, A2)	2. B(T5) → UNDO = {T2,T3,T5}. REDO = {T4} Setup
B(T3)	3. C(T5) → UNDO = {T2,T3}. REDO = {T4, T5}
C(T1)	<hr/>
B(T4)	4. D(O6)
7. U(T3,O2,B3,A3)	5. O5 = B7
9. U(T4,O3,B4,A4)	6. O3 = B5
CK(T2,T3,T4)	7. O2 = B3
1. C(T4)	8. O1=B1
2. B(T5)	
6. U(T3,O3,B5,A5)	
10. U(T5,O4,B6,A6)	
5. D(T3,O5,B7)	
A(T3)	
3. C(T5)	
4. I(T2,O6,A8)	

Undo

12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

40

## 4. Fase REDO

B(T1)	0. UNDO = {T2,T3,T4}. REDO = {}	
B(T2)		
8. U(T2, O1, B1, A1)	1. C(T4) → UNDO = {T2, T3}. REDO = {T4}	
I(T1, O2, A2)	2. B(T5) → UNDO = {T2,T3,T5}. REDO = {T4}	Setup
B(T3)		
C(T1)	3. C(T5) → UNDO = {T2,T3}. REDO = {T4, T5}	
B(T4)		
7. U(T3,O2,B3,A3)	4. D(O6)	
9. U(T4,O3,B4,A4)	5. O5 = B7	
CK(T2,T3,T4)		
1. C(T4)	6. O3 = B5	Undo
2. B(T5)		
6. U(T3,O3,B5,A5)	7. O2 = B3	
10. U(T5,O4,B6,A6)	8. O1=B1	
5. D(T3,O5,B7)		
A(T3)	9. O3 = A4	
3. C(T5)		Redo
4. I(T2,O6,A8)	10. O4 = A6	

12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

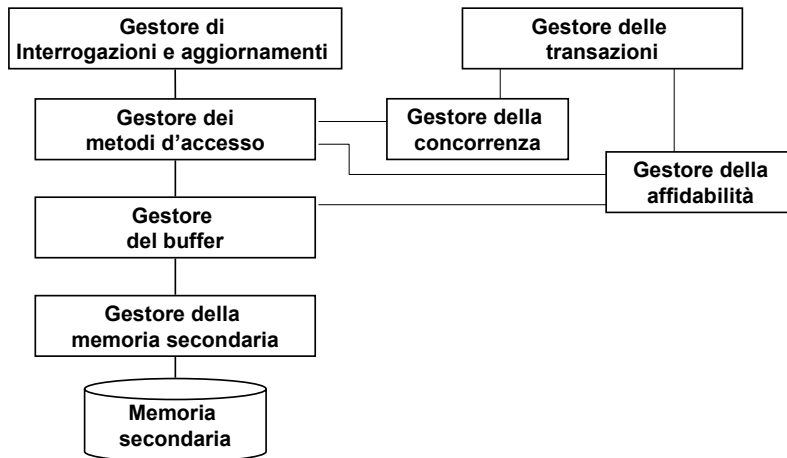
41

## Ripresa a freddo

- Si ripristinano i dati a partire dal backup
- Si eseguono le operazioni registrate sul giornale fino all'istante del guasto
- Si esegue una ripresa a caldo

## Gestore degli accessi e delle interrogazioni

## Gestore delle transazioni



12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

43

## Controllo di concorrenza

- La concorrenza è fondamentale: decine o centinaia di transazioni al secondo, non possono essere seriali
- Esempi: banche, prenotazioni aeree

### Modello di riferimento

- Operazioni di input-output su oggetti astratti  $x, y, z$

### Problema

- Anomalie causate dall'esecuzione concorrente, che quindi va governata

12/10/2003

Basi di dati, vol.2 cap.2 Gestione delle transazioni

44

## Perdita di aggiornamento

- Due transazioni identiche:
  - $t_1 : r(x), x = x + 1, w(x)$
  - $t_2 : r(x), x = x + 1, w(x)$
- Inizialmente  $x=2$ ; dopo un'esecuzione seriale  $x=4$
- Un'esecuzione concorrente:

$t_1$	$t_2$
bot	
$r_1(x)$	
$x = x + 1$	
	bot
	$r_2(x)$
	$x = x + 1$
$w_1(x)$	
commit	
	$w_2(x)$
	commit

- Un aggiornamento viene perso:  $x=3$

## Letture sporca

$t_1$	$t_2$
bot	
$r_1(x)$	
$x = x + 1$	
$w_1(x)$	
	bot
	$r_2(x)$
abort	
	commit

- Aspetto critico:  $t_2$  ha letto uno stato intermedio ("sporco") e lo può comunicare all'esterno

## Letture inconsistenti

- $t_1$  legge due volte:

$t_1$	$t_2$
bot	
$r_1(x)$	
	bot
	$r_2(x)$
	$x = x + 1$
	$w_2(x)$
	commit
$r_1(x)$	
commit	

- $t_1$  legge due valori diversi per  $x$  !

## Aggiornamento fantasma

- Assumere ci sia un vincolo  $y + z = 1000$ ;

$t_1$	$t_2$
bot	
$r_1(y)$	
	bot
	$r_2(y)$
	$y = y - 100$
	$r_2(z)$
	$z = z + 100$
	$w_2(y)$
	$w_2(z)$
	commit
$r_1(z)$	
$s = y + z$	
commit	

- $s = 1100$ : il vincolo sembra non soddisfatto,  $t_1$  vede un aggiornamento non coerente



## Inserimento fantasma

$t_1$

bot  
"legge gli stipendi degli impiegati  
del dip A e calcola la media"

"legge gli stipendi degli impiegati  
del dip A e calcola la media"  
commit

$t_2$

bot  
"inserisce un impiegato in A"  
commit

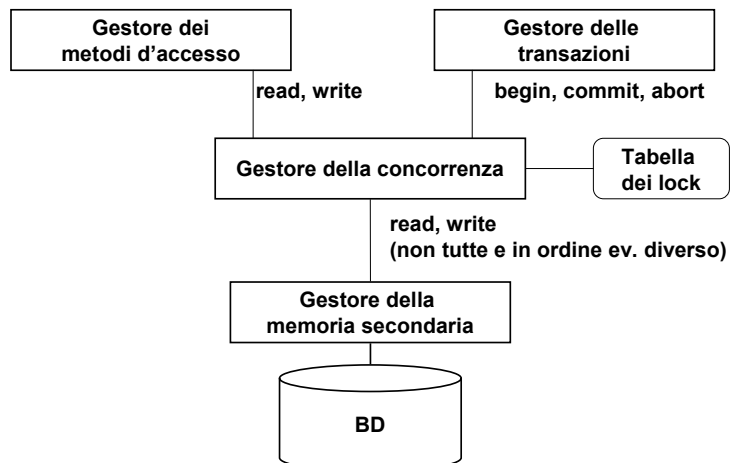
## Anomalie

- Perdita di aggiornamento W-W
- Lettura sporca R-W (o W-W) con abort
- Letture inconsistenti R-W
- Aggiornamento fantasma R-W
- Inserimento fantasma R-W su dato "nuovo"

## Concurrency control theory

- Transaction: sequence of read or write actions (no additional semantics)
- Each transaction has a unique, system-assigned transaction identifier
- Each transactions is initiated by the **start transaction** command and terminated by **commit** o **abort**
- Example:  $t_1 : r_1(x) r_1(y) w_1(x) w_1(y)$

## Gestore della concorrenza (ignorando buffer e affidabilità)



## Schedule

- Sequenza di operazioni di input/output operations di transazioni concorrenti
- Esempio:

$$S_1 : r_1(x) r_2(z) w_1(x) w_2(z)$$

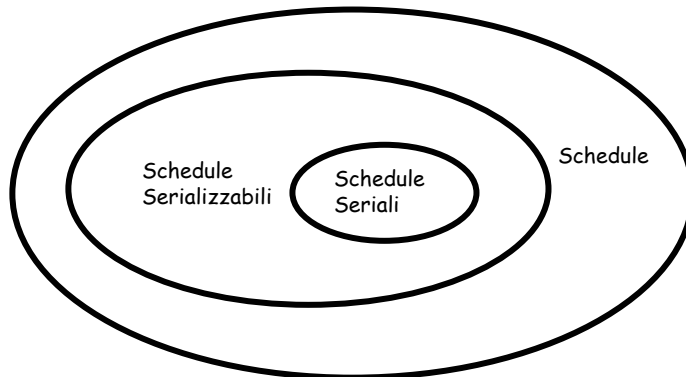
- Ipotesi semplificativa (che rinvieremo in futuro, in quanto non accettabile in pratica):
  - consideriamo la **commit-proiezione** e ignoriamo le transazioni che vanno in abort, rimuovendo tutte le loro azioni dallo schedule

## Controllo di concorrenza

- *Obiettivo*: evitare le anomalie
- *Scheduler*: un sistema che accetta o rifiuta (o riordina) le operazioni richieste dalle transazioni
- *Schedule seriale*: le transazioni sono separate, una alla volta
$$S_2 : r_0(x) r_0(y) w_0(x) r_1(y) r_1(x) w_1(y) r_2(x) r_2(y) r_2(z) w_2(z)$$
- *Schedule serializzabile*: produce lo stesso risultato di uno schedule seriale sulle stesse transazioni
  - Richiede una nozione di equivalenza fra schedule

## Idea base

- Individuare classi di schedule serializzabili che siano sottoclassi degli schedule possibili, siano serializzabili e la cui proprietà di serializzabilità sia verificabile a costo basso



## View-Serializzabilità

- Definizioni preliminari:
  - $r_i(x)$  **legge-da**  $w_j(x)$  in uno schedule  $S$  se  $w_j(x)$  precede  $r_i(x)$  in  $S$  e non c'è  $w_k(x)$  fra  $r_i(x)$  e  $w_j(x)$  in  $S$
  - $w_i(x)$  in uno schedule  $S$  è **scrittura finale** se è l'ultima scrittura dell'oggetto  $x$  in  $S$
- Schedule **view-equivalenti** ( $S_i \approx_v S_j$ ): hanno la stessa relazione **legge-da** e le stesse scritture finali
- Uno schedule è **view-serializzabile** se è view-equivalente ad un qualche schedule seriale
- L'insieme degli schedule view-serializzabili è indicato con **VSR**

## View serializzabilità: esempi

- $S_3 : w_0(x) r_2(x) r_1(x) w_2(x) w_2(z)$   
 $S_4 : w_0(x) r_1(x) r_2(x) w_2(x) w_2(z)$   
 $S_5 : w_0(x) r_1(x) w_1(x) r_2(x) w_1(z)$   
 $S_6 : w_0(x) r_1(x) w_1(x) w_1(z) r_2(x)$ 
  - $S_3$  è view-equivalente allo schedule seriale  $S_4$  (e quindi è view-serializzabile)
  - $S_5$  è view-equivalente sonoa  $S_4$ , ma è view-equivalente allo schedule seriale  $S_6$ , e quindi è view-serializzabile
- $S_7 : r_1(x) r_2(x) w_1(x) w_2(x)$  (perdita di aggiornamento)  
 $S_8 : r_1(x) r_2(x) w_2(x) r_1(x)$  (letture inconsistenti)  
 $S_9 : r_1(x) r_1(y) r_2(z) r_2(y) w_2(y) w_2(z) r_1(z)$  (aggiornamento fantasma)
  - $S_7, S_8, S_9$  non view-serializzabili

## View serializzabilità

- Complessità:
  - la verifica della view-equivalenza di due dati schedule:
    - polinomiale
  - decidere sulla View serializzabilità di uno schedule:
    - problema NP-completo
- Non è utilizzabile in pratica

## Conflict-serializzabilità

- Definizione preliminare:
  - Un'azione  $a_i$  è in *confitto* con  $a_j$  ( $i \neq j$ ), se operano sullo stesso oggetto e almeno una di esse è una scrittura. Due casi:
    - conflitto *read-write* ( $rw$  o  $wr$ )
    - conflitto *write-write* ( $ww$ ).
- *Schedule conflict-equivalenti* ( $S_i \approx_C S_j$ ): includono le stesse operazioni e ogni coppia di operazioni in conflitto compare nello stesso ordine in entrambi
- Uno schedule è *conflict-serializable* se è view-equivalente ad un qualche schedule seriale
- L'insieme degli schedule conflict-serializzabili è indicato con **CSR**

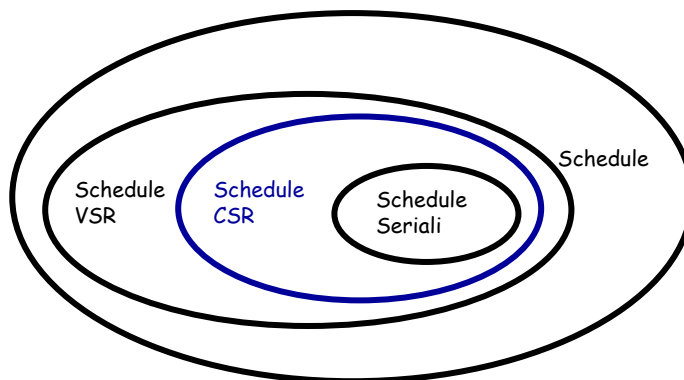
## CSR e VSR

- Ogni schedule conflict-serializzabile è view-serializzabile, ma non necessariamente viceversa
- *Controesempio per la non necessità:*
- $r1(x) w2(x) w1(x) w3(x)$ 
  - view-serializzabile: view-equivalente a  $r1(x) w1(x) w2(x) w3(x)$
  - non conflict-serializzabile
- *Sufficienza:* vediamo

## CSR implica VSR

- CSR: esiste schedule seriale conflict-equivalente
- VSR: esiste schedule seriale view-equivalente
- Per dimostrare che CSR implica VSR è sufficiente dimostrare che la conflict-equivalenza  $\approx_C$  implica la view-equivalenza  $\approx_V$ , cioè che se due schedule sono  $\approx_C$  allora sono  $\approx_V$
- Quindi, supponiamo  $S_1 \approx_C S_2$  e dimostriamo che  $S_1 \approx_V S_2$   
I due schedule hanno:
  - stesse scritture finali: se così non fosse, ci sarebbero almeno due scritture in ordine diverso e poiché due scritture sono in conflitto i due schedule non sarebbero  $\approx_C$
  - stessa relazione “legge-da”: se così non fosse, ci sarebbero scritture in ordine diverso o coppie lettura-scrittura in ordine diverso e quindi, come sopra sarebbe violata la  $\approx_C$

## CSR e VSR



## Verifica di conflict-serializzabilità

- Per mezzo del **grafo dei conflitti**:
  - un nodo per ogni transazione  $t_i$
  - un arco (orientato) da  $t_i$  a  $t_j$  se c'è almeno un conflitto fra un'azione  $a_i$  e un'azione  $a_j$  tale che  $a_i$  precede  $a_j$
- Teorema
  - **Uno schedule è in CSR se e solo se il grafo è aciclico**

## CSR e aciclicità del grafo dei conflitti

- Se uno schedule  $S$  è CSR allora è  $\approx_C$  ad uno schedule seriale. Supponiamo le transazioni nello schedule seriale ordinate secondo il TID:  $t_1, t_2, \dots, t_n$ . Poiché lo schedule seriale ha tutti i conflitti nello stesso ordine dello schedule  $S$ , nel grafo di  $S$  ci possono essere solo archi  $(i,j)$  con  $i < j$  e quindi il grafo non può avere cicli, perché un ciclo richiede almeno un arco  $(i,j)$  con  $i > j$ .
- Se il grafo di  $S$  è aciclico, allora esiste fra i nodi un "ordinamento topologico" (cioè una numerazione dei nodi tale che il grafo contiene solo archi  $(i,j)$  con  $i < j$ ). Lo schedule seriale le cui transazioni sono ordinate secondo l'ordinamento topologico è equivalente a  $S$ , perché per tutti i conflitti  $(i,j)$  si ha sempre  $i < j$ .



## Controllo della concorrenza in pratica

- Anche la conflict-serializzabilità, pur più rapidamente verificabile (l'algoritmo, con opportune strutture dati richiede tempo lineare), è inutilizzabile in pratica
- La tecnica sarebbe efficiente se potessimo conoscere il grafo dall'inizio, ma così non è: uno scheduler deve operare "incrementalmente", cioè ad ogni richiesta di operazione decidere se eseguirla subito oppure fare qualcos'altro; non è praticabile mantenere il grafo, aggiornarlo e verificarne l'aciclicità ad ogni richiesta di operazione
- Inoltre, la tecnica si basa sull'ipotesi di commit-proiezione
- In pratica, si utilizzano tecniche che
  - garantiscono la conflict-serializzabilità senza dover costruire il grafo
  - non richiedono l'ipotesi della commit-proiezione

## Lock

- Principio:
  - Tutte le letture sono precedute da *r\_lock* (lock condiviso) e seguite da *unlock*
  - Tutte le scritture sono precedute da *w\_lock* (lock esclusivo) e seguite da *unlock*
- Quando una transazione prima legge e poi scrive un oggetto, può:
  - richiedere subito un lock esclusivo
  - chiedere prima un lock condiviso e poi uno esclusivo (*lock escalation*)
- Il *lock manager* riceve queste richieste dalle transazioni e le accoglie o rifiuta, sulla base della tavola dei conflitti

## Gestione dei lock

- Basata sulla tavola dei conflitti

Riochiesta	<i>free</i>	Stato della risorsa	<i>r_locked</i>	<i>w_locked</i>
<i>r_lock</i>	OK / <i>r_locked</i>	OK / <i>r_locked</i>	NO / <i>w_locked</i>	
<i>w_lock</i>	OK / <i>w_locked</i>	NO / <i>r_locked</i>	NO / <i>w_locked</i>	
<i>unlock</i>	error	OK / depends	OK / free	

- Un contatore tiene conto del numero di "lettori"; la risorsa è rilasciata quando il contatore scende a zero
- Se la risorsa non è concessa, la transazione richiedente è posta in attesa (eventualmente in coda), fino a quando la risorsa non diventa disponibile
- Il lock manager gestisce una tabella dei lock, per ricordare la situazione

## Locking a due fasi

- Usato da quasi tutti i sistemi
- Garantisce "a priori" la conflict-serializzabilità a-priori
- Basata su due regole:
  - "proteggere" tutte le letture e scritture con lock
  - un vincolo sulle richieste e i rilasci dei lock:
    - **una transazione, dopo aver rilasciato un lock, non può acquisirne altri**

## 2PL e CSR

- Ogni schedule 2PL e' anche conflict serializzabile, ma non necessariamente viceversa
- Controesempio per la non necessita':

$r_1(x) w_1(x) r_2(x) w_2(x) r_3(y) w_1(y)$

- Viola 2PL
  - Conflict-serializzabile
- Sufficienza: vediamo

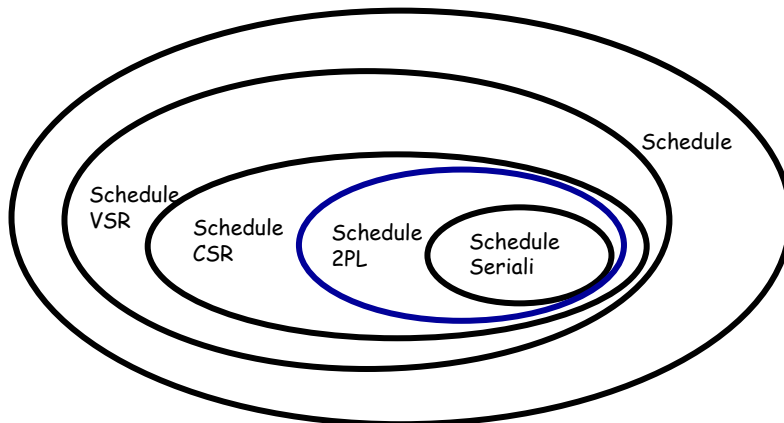
## 2PL implica CSR

- S schedule 2PL
- Consideriamo per ciascuna transazione l'istante in cui ha tutte le risorse e sta per rilasciare la prima
- Ordiniamo le transazioni in accordo con questo valore temporale e consideriamo lo schedule seriale corrispondente
- Vogliamo dimostrare che tale schedule è equivalente ad S:
  - allo scopo, consideriamo un conflitto fra un'azione di  $t_i$  e un'azione dei  $t_j$  con  $i < j$ ; è possibile che compaiano in ordine invertito in S? no, perché in tal caso  $t_j$  dovrebbe aver rilasciato la risorsa in questione prima della sua acquisizione da parte di  $t_i$

## Locking a due fasi stretto

- Condizione aggiuntiva:
  - **I lock possono essere rilasciati solo dopo il commit o abort**
- Supera la necessità dell'ipotesi di commit-proiezione (ed elimina il rischio di letture sporche)

## CSR, VSR e 2PL



## Controllo di concorrenza basato su timestamp

- Tecnica alternativa al 2pL
- **Timestamp:**
  - identificatore che definisce un ordinamento totale sugli eventi di un sistema
- Ogni transazione ha un timestamp che rappresenta l'istante di inizio della transazione
- Uno schedule è accettato solo se riflette l'ordinamento seriale delle transazioni indotto dai timestamp

## Dettagli

- Lo scheduler ha due contatori  $RTM(x)$  e  $WTM(x)$  per ogni oggetto
- Lo scheduler riceve richieste di letture e scritture (con indicato il timestamp della transazione):
  - $read(x,ts)$ :
    - se  $ts < WTM(x)$  allora la richiesta è respinta e la transazione viene uccisa;
    - altrimenti, la richiesta viene accolta e  $RTM(x)$  è posto uguale al maggiore fra  $RTM(x)$  e  $ts$
  - $write(x,ts)$ :
    - se  $ts < WTM(x)$  o  $ts < RTM(x)$  allora la richiesta è respinta e la transazione viene uccisa,
    - altrimenti, la richiesta e  $WTM(x)$  è posto uguale a  $ts$
- Vengono uccise molte transazioni
- Per funzionare anche senza ipotesi di commit-proiezione, deve "bufferizzare" le scritture fino al commit (con attese)

## Esempio

$RTM(x) = 7$

$WTM(x) = 4$

Richiesta	Risposta	Nuovo valore
<i>read(x,6)</i>	ok	
<i>read(x,8)</i>	ok	$RTM(x) = 8$
<i>read(x,9)</i>	ok	$RTM(x) = 9$
<i>write(x,8)</i>	no, $t_8$ uccisa	
<i>write(x,11)</i>	ok	$WTM(x) = 11$
<i>read(x,10)</i>	no, $t_{10}$ uccisa	

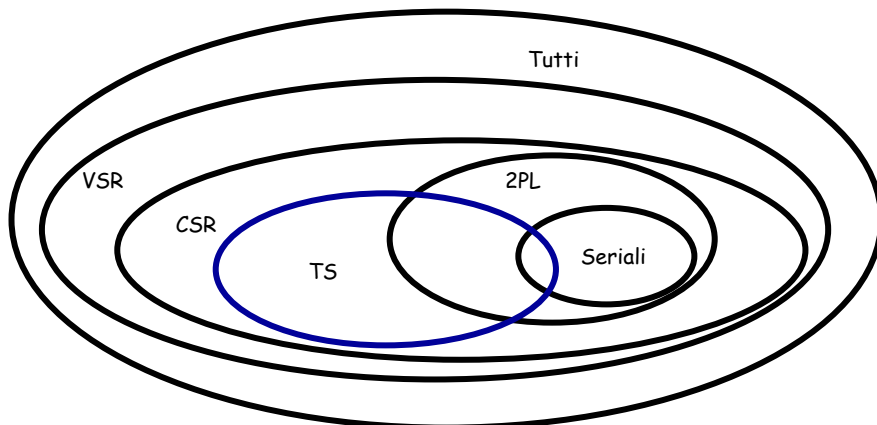
## Multiversion concurrency control

- Main idea: writes generate new copies, reads make access to the correct copy
- Writes generate new copies each with a WTM. At any time,  $N > 1$  copies of each object  $x$  are active, with  $WTM_N(x)$ . There is only one global  $RTM(x)$
- Mechanism:
  - *read(x,ts)*: is always accepted.  $x_k$  selected for reading such that: if  $ts > WTM_N(x)$ , then  $k = N$ , otherwise  $k$  is taken such that  $WTM_k(x) < ts < WTM_{k+1}(x)$
  - *write(x,ts)*: if  $ts < RTM(x)$  the request is refused, otherwise a new version of the item of data is added ( $N$  increased by one) with  $WTM_N(x) = ts$
- Old copies are discarded when there are no read transactions interested in their values

## 2PL vs TS

- Sono incomparabili
  - Schedule in TS ma non in 2PL  
 $r_1(x) w_1(x) r_2(x) w_2(x) r_0(y) w_1(y)$
  - Schedule in 2PL ma non in TS  
 $r_2(x) w_2(x) r_1(x) w_1(x)$
  - Schedule in TS e in 2PL  
 $r_1(x) r_2(y) w_2(y) w_1(x) r_2(x) w_2(x)$

## CSR, VSR, 2PL e TS



## 2PL vs TS

- In 2PL le transazioni sono poste in attesa In TS uccise e rilanciate
- Per rimuovere la commit proiezione, attesa per il commit in entrambi i casi
- 2PL può causare deadlock (vedremo)
- Le ripartenze sono di solito più costose delle attese:
  - conviene il 2PL

## Lock management

- Interface:
  - `r_lock(T, x, errcode, timeout)`
  - `w_lock(T, x, errcode, timeout)`
  - `unlock(T, x)`

T: transaction identifier  
X: data element  
timeout: max wait in queue
- If timeout expires, `errcode` signals an error, typically the transaction rolls back and restarts



## Hierarchical locking

- In many real systems locks can be specified at different granularity, e.g. tables, fragments, tuples, fields. These are organized in a hierarchy (possibly a directed acyclic graph)
- 5 locking modes:
  - 2 are shared and exclusive, renamed as:
    - XL: exclusive lock
    - SL: shared lock
  - 3 are new:
    - ISL: intention shared lock
    - IXL: intention exclusive lock
    - SIXL: shared intention-exclusive lock
- The choice of lock granularity is left to application designers;
  - too coarse: many resources are blocked
  - too fine: many locks are requested

## Hierarchical locking protocol

- Locks are requested from the root to descendents in a hierarchy
- Locks are released starting at the node locked and moving up the tree
- In order to request an SL or ISL on a node, a transaction must already hold an ISL or IXL lock on the parent node
- In order to request an IXL, XL, or SIXL on a node, a transaction must already hold an SIXL or IXL lock on the parent node
- The new conflict table is shown in the next slide

## Conflicts in hierarchical locking

Request	Resource state				
	ISL	IXL	SL	SIXL	XL
ISL	OK	OK	OK	OK	No
IXL	OK	OK	No	No	No
SL	OK	No	OK	No	No
SIXL	OK	No	No	No	No
XL	No	No	No	No	No

## Stallo (deadlock)

- Attese incrociate: due transazioni detengono ciascuna una risorsa e aspettano la risorsa detenuta dall'altra
- Esempio:
  - $t_1$ :  $read(x)$ ,  $write(y)$
  - $t_2$ :  $read(y)$ ,  $write(x)$
  - Schedule:  
 $r\_lock_1(x)$ ,  $r\_lock_2(y)$ ,  $read_1(x)$ ,  $read_2(y)$   $w\_lock_1(y)$ ,  
 $w\_lock_2(x)$

## Risoluzione dello stallo

- Uno stallo corrisponde ad un ciclo nel grafo delle attese (nodo=transazione, arco=attesa)
- Tre tecniche
  1. Timeout (problema: scelta dell'intervallo, con trade-off)
  2. Rilevamento dello stallo
  3. Prevenzione dello stallo
- Rilevamento: ricerca di cicli nel grafo delle attese
- Prevenzione: uccisione di transazioni "sospette" (può esagerare)

## Livelli di isolamento in SQL:1999 (e JDBC)

- Le transazioni possono essere definite **read-only** (non possono richiedere lock esclusivi)
- Il livello di isolamento può essere scelto per ogni transazione
  - **read uncommitted** permette letture sporche, letture inconsistenti, aggiornamenti fantasma e inserimenti fantasma
  - **read committed** evita letture sporche ma permette letture inconsistenti, aggiornamenti fantasma e inserimenti fantasma
  - **repeatable read** evita tutte le anomalie esclusi gli inserimenti fantasma
  - **serializable** evita tutte le anomalie
- Nota:
  - la perdita di aggiornamento è sempre evitata

## Livelli di isolamento: implementazione

- Sulle scritture si ha sempre il 2PL stretto (e quindi si evita la perdita di aggiornamento)
- **read uncommitted:**
  - nessun lock in lettura (e non rispetta i lock altrui)
- **read committed:**
  - lock in lettura (e rispetta quelli altrui), ma senza 2PL
- **repeatable read:**
  - 2PL anche in lettura, con lock sui dati
- **serializable:**
  - 2PL con lock di predicato

## Lock di predicato

- Caso peggiore:
  - sull'intera relazione
- Se siamo fortunati:
  - sull'indice

## Casi di studio per il tuning di basi di dati

dal testo:

D. Shasha.  
Database Tuning: a principled approach.  
Prentice-Hall, 1992

## Caso 1

- Abbiamo una base di dati con dieci relazioni su due dischi: cinque e cinque; su un disco c'è anche il log.
- Compriamo un nuovo disco; che cosa ci mettiamo?

### Possibile risposta

- Il log

## Caso 2

- Il tempo di risposta è variabile, in particolare ci sono rallentamenti quando vengono eseguite operazioni DDL

### Possibile motivazione

- Le operazioni DDL richiedono lock in scrittura sul catalogo

## Caso 3

- Transazioni così organizzate sono insoddisfacenti:
  - attribuzione di un nuovo numero di pratica
  - richiesta di informazioni interattive
  - effettuazione dell'inserimento

### Possibile soluzione

- transazione troppo lunga; vanno riordinati i passi, e nella transazione ci devono essere solo il primo e il terzo

## Caso 4

- Una transazione così organizzata eseguita a fine mese, di sera, è inefficiente:  
per ogni conto stampare tutte le info ...

### Possibile soluzione

- essendo di sola lettura, di sera (tempo morto) potrebbe essere senza lock (Read Uncommitted)

## Casi 5 e 6

- Nell'ambito di una transazione, si calcola lo stipendio medio per ciascun dipartimento. Contemporaneamente si fanno modifiche su singoli stipendi.
- Le prestazioni sono insoddisfacenti.

### Possibili soluzioni

- si può eseguire in un tempo morto (senza aggiornamenti) senza lock (Read Uncommitted)
- se sono tollerate (leggere) inconsistenze, si può procedere senza lock (Read Uncommitted)
- si può fare una copia e lavorare su di essa (dati non attuali)
- se nessuna delle alternative è praticabile (non ci sono tempi morti e si vogliono dati attuali e consistenti) si può provare con Repeatable Read (non c'è rischio di "phantom")

## Caso 7

- Un'applicazione prevede:
  - migliaia di inserimenti ogni ora
  - centinaia di migliaia di piccoli aggiornamenti ogni ora
- Gli inserimenti arrivano in transazioni grandi ogni mezz'ora e durano 5 minuti. In queste fasi le prestazioni sono inaccettabili (tempo di risposta 30 sec, rispetto a mezzo secondo)
- **Possibili soluzioni**
- spezzare le transazioni con gli inserimenti
- riorganizzare i file evitando strutture ordinate

## Caso 8

- Un'applicazione che prevede un'istruzione SQL all'interno di un ciclo è lenta (e usa molto tempo di CPU)
- **Possibile soluzione**
- usare bene il cursore (facciamo fare i cicli e soprattutto i join all'SQL)



## Caso 9

- Un disco è inefficiente anche se non pieno. Ci sono relativamente pochi inserimenti e molte letture lunghe (scansioni)

### Possibili soluzioni

- separare il log su un disco a sé
- riorganizzare i file in modo contiguo

## Caso 10

- Una società di servizi emette tutte le bollette a fine mese, con un programma che ha bisogno di tutta la notte, impedendo così l'esecuzione di altri programmi batch che sarebbero necessari

### Possibili soluzioni

- è proprio necessario che le bollette siano emesse tutte insieme?
- se è proprio necessario, magari facciamolo durante il week-end (tempo morto più lungo)