Kapitel 16: Transaktionsorientierte Daten-Recovery

16.1 Crash-Recovery

Fehlerkategorien (die die Atomarität oder Persistenz von Transaktionen gefährden)**:**

1) Fehler im Anwendungsprogramm
2) Fehler in der Systemsoftware (BS oder DBS),
   die zum "Systemabsturz" (engl.: system failure, (soft) crash) führen
   - "Bohrbugs": deterministisch und reproduzierbar;
     sind in gut ausgetester Systemsoftware so gut wie eliminiert
   - "Heisenbugs": schwer einzugrenzen und praktisch nicht reproduzierbar;
     treten vor allem in Überlastsituationen im Mehrbenutzerbetrieb auf
3) Stromausfall
4) Transienter Hardwarefehler (CPU, Speicher)
5) Fehlbedienung (Systemadministrator, Operateur, Techniker)
6) Plattenfehler (Media-Fehler),
   der zum Verlust von permanent gespeicherten Daten führt
   (engl.: disk failure, hard crash)
7) Katastrophe (Feuer, Erdbeben, etc.)

Fehlerbehandlung durch das DBS:

1) Unvollendete Transaktionen von fehlerhaft beendeten Programmen werden zurückgesetzt.
2) - 5)
   Das DBS wird im *Warmstart* neu gestartet und führt dabei eine ***Crash-Recovery*** durch.
   Dabei werden alle vor dem Absturz unvollendeten Transaktionen zurückgesetzt
   (***Undo*** von Transaktionen).
   Änderungen von beendeten Transaktionen werden nötigenfalls DBS-intern wiederholt
   (***Redo*** von Transaktionen).
   Bemerkungen:
   - Transaktionen, für die die Recovery ein Undo durchgeführt hat, müssen vom Anwendungsprogramm oder sogar vom Endbenutzer neu gestartet werden.
   - Das Prinzip einer "Backward Recovery" (Undo mit anschließender Fortsetzung der Verarbeitung) hat sich gegenüber einer "Forward Recovery" (z.B. instruktionssynchrone Verarbeitung auf einem zweiten Rechner, der im Fehlerfall die Verarbeitung fortsetzt) vor allem im Hinblick auf "Heisenbugs" sehr bewährt.
   - Die Fehlerkategorie 3 kann durch ununterbrechbare Stromversorgung eliminiert werden.
6) Wiederherstellung der verlorenen Daten durch Aufsetzen auf einer Archivkopie der Datenbank und Wiederholung der Änderungen abgeschlossener Transaktionen (***Media-Recovery***).
7) Änderungen abgeschlossener Transaktionen müssen auf einem zweiten, genügend weit entfernten Rechner doppelt erfaßt werden; im Katastrophenfall übernimmt der zweite Rechner die Verarbeitung.

Arbeitsprinzipien der Recovery-Komponente:

- Um im Fehlerfall Undo und Redo durchführen zu können, müssen Änderungen auf einem hinreichend stabilen Speichermedium (für die Crash-Recovery auf Platte, für die Media-Recovery auf mehreren Platten und/oder Band) protokolliert werden. Für jede Änderung gibt es eine *Undo-Information*, um die Änderung rückgängig machen zu können, und eine *Redo-Information*, um die Änderung wiederholen zu können.

- ***Atomaritätsprinzip:***
  Vor einer Veränderung der permanenten Datenbank (durch Zurückschreiben geänderter Seiten aus dem Puffer auf die Platte) muß die entsprechende Undo-Information auf stabilen Speicher geschrieben werden.

- ***Persistenzprinzip:***
  Vor dem Commit einer Transaktion muß die entsprechende Redo-Information auf stabilen Speicher geschrieben werden.

- ***Idempotenzprinzip:***
  Da es während eines Warmstarts zu einem erneuten Ausfall kommen kann, müssen Recovery-Maßnahmen beliebig oft wiederholbar sein.
  Dabei muß sichergestellt sein, daß
  - zweimaliges Undo derselben Änderung denselben Effekt hat wie einmaliges Undo,
  - zweimaliges Redo derselben Änderung denselben Effekt hat wie einmaliges Redo,
  - Undo einer Änderung, die durch den Systemabsturz bereits verloren gegangen ist, keinen Effekt hat und
  - Redo einer Änderung, die durch den Systemabsturz gar nicht verloren gegangen ist, keinen Effekt hat.

Generelle Zielsetzungen der Recovery-Komponente:

- Hohe Verfügbarkeit = MTTF / (MTTF + MTTR)

  (Eine Verfügbarkeit von 99 % impliziert beispielsweise eine "Auszeit" von ca. 5000 Minuten pro Jahr, während der das System nicht verfügbar ist;
  eine Verfügbarkeit von 99.9 % impliziert eine "Auszeit" von ca. 500 Minuten pro Jahr.)

  mit MTTF = Mean Time To Failure

  MTTR = Mean Time to Repair

- Schnelle Recovery beim Warmstart (kurze MTTR)
- Geringer Overhead im laufenden Normalbetrieb

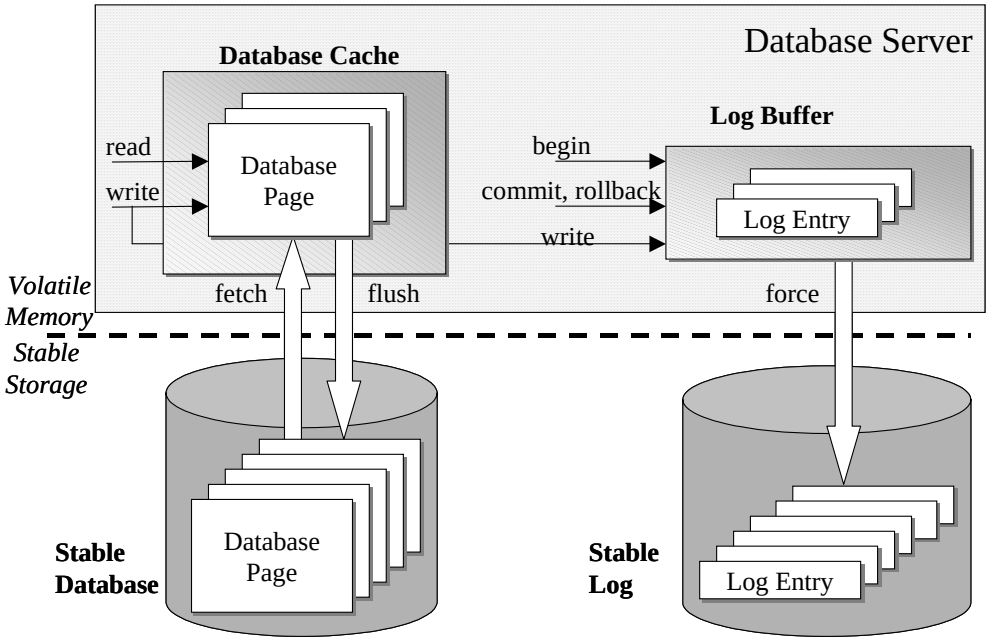
Implementierung der Crash-Recovery mittels Logging:

- Änderungen werden in Form von ***Logsätzen*** in einer Logdatei protokolliert.
  Um nicht für alle Änderung einen Plattenzugriff durchführen zu müssen, werden Logsätze zunächst in einen ***Logpuffer*** im Hauptspeicher geschrieben und nur bei Bedarf (s.u.) sequentiell in die ***Logdatei*** auf Platte geschrieben.

  Bemerkung:
  Um sicherzustellen, daß für die Logdatei tatsächlich immer sequentielle I/Os stattfinden, sollte die Logdatei auf einer separaten Platte liegen.
- Ein Logsatz enthält in der Regel sowohl die ***Undo-Information*** als auch die ***Redo-Information*** einer Änderung.
  Im einfachsten Fall bezieht sich ein Logsatz bzw. die darin protokollierte Änderung auf genau eine Seite der Datenbank. Der Logsatz beschreibt entweder den Zustand der Seite bzw. der geänderten Bytes vor und nach der Änderung (physisches Logging, zustandsorientiertes Logging) oder eine Operation auf der Byte-Ebene (physiologisches Logging, operationsorientiertes Logging auf der Byte-Ebene).
  Ein Spezialfall des zustandsorientierten Logging ist das Protokollieren ganzer Seiten jeweils vor und nach einer Änderung ("Before-Images" und "After-Images").

  Bemerkung:
  Operationsorientiertes Logging ist auch für komplexere Operationen möglich, die mehrere Seiten ändern (logisches Logging). Diese Variante ist aber schwieriger zu realisieren und daher in kommerziellen Systemen selten zu finden.
- Zusätzlich werden Logsätze für die Beendigung von Transaktionen (EOT und RBT) geschrieben, sowie optional auch für den Beginn von Transaktionen (BOT).
- Logsätze sind - über Logpuffer und Logdatei hinweg - transaktionsweise rückwärts verkettet. Dadurch können einzelne Transaktionen auch im laufenden Normalbetrieb einfach zurückgesetzt werden.
- Logsätze sind fortlaufend numeriert. Die Nummer eines Logsatzes - genannt ***"Log Sequence Number" (LSN, LogSeqNo)*** - gibt den relativen Zeitpunkt einer Änderung an. Sie kann auch - je nach Implementierung - zur Adressierung der Logsätze in der Logdatei dienen.
  Die LSN der jeweils jüngsten Änderung einer Seite wird in einem speziellen Feld ***"SeqNo" im Seiten-Header*** gespeichert.
- Der Logpuffer muß in den folgenden Fällen in die Logdatei auf Platte geschrieben werden:
  - ***Undo-Regel:*** vor dem Zurückschreiben einer geänderten Seite aus dem Puffer in die Datenbank auf der Platte, sofern die Seite Änderungen von noch nicht abgeschlossenen Transaktionen enthält (*Write Ahead Logging, WAL-Prinzip*)
  - ***Redo-Regel:*** unmittelbar vor dem Commit einer Transaktion, die Änderungen durchgeführt hat.

Architektur für das Logging im laufenden Normalbetrieb:



Vorgehen beim Warmstart nach einem Crash:

1) ***Analysephase:***
   Die Logdatei wird sequentiell gelesen (in aufsteigender LSN-Reihenfolge), und dabei wird bestimmt, welche Transaktionen vor dem Crash beendet wurden ("Gewinner") und welche unvollendet waren ("Verlierer").

2) ***Redo-Phase:***
   Die Logdatei wird nochmals sequentiell gelesen (in aufsteigender LSN-Reihenfolge), und dabei werden die Änderungen der Gewinner – oder beim sog. *Redo-History-Algorithmus* – alle Änderungen wiederholt, sofern die Änderungen noch nicht in der Datenbank enthalten sind. Dazu wird getestet, ob die LSN des Logsatzes größer (d.h. jünger) ist als die LSN im Header der betreffenden Seite; und nur wenn dies der Fall ist, wird die Änderung wiederholt, wobei die LSN im Header der Seite entsprechend erhöht wird. Dieser Test stellt die Idempotenz der Redo-Phase sicher.

3) ***Undo-Phase:***
   Die Logdatei wird nochmals sequentiell gelesen (in absteigender LSN-Reihenfolge), und dabei werden die Änderungen der Verlierer rückgängig gemacht, sofern die Änderungen in der Datenbank enthalten sind. Dazu wird getestet, ob die LSN des Logsatzes kleiner oder gleich der LSN im Header der betreffenden Seite ist (d.h. nicht jünger); und nur wenn dies der Fall ist, wird die Änderung rückgängig gemacht und die LSN im Header der Seite erniedrigt.
   Dieser Test stellt die Idempotenz der Undo-Phase sicher.

Verbesserung der Effizienz:

- *im laufenden Normalbetrieb:*
  Wegen des Persistenzprinzips muß eigentlich bei jedem Commit einer Änderungstransaktion der Logpuffer in die Logdatei auf Platte geschrieben werden. Man kann die Anzahl der Log-I/Os reduzieren, indem man das Schreiben mehrerer EOT-Logsätze zusammenfaßt. Dazu muß das Commit von Transaktionen (genauer: das Melden des SQLCODE 0 für die COMMIT-WORK-Anweisung des Anwendungsprogramms) ggf. verzögert werden, bis der Logpuffer voll ist, der Logpuffer wegen des WAL-Prinzips auf Platte geschrieben werden muß oder ein Commit bereits sehr lange verzögert wurde. Diese Technik nennt man *Group-Commit*.

- *beim Warmstart:*
  Die Analysephase und die Redo-Phase der Recovery müssen im Prinzip die gesamte Logdatei lesen, obwohl vermutlich nur der jüngere Teil der Logdatei relevant ist.
  Dies kann verbessert werden, indem man im laufenden Normalbetrieb periodisch *Sicherungspunkte (Checkpoints)* erzeugt, die dazu beitragen, den Analyse- und Redo-Aufwand beim Warmstart zu begrenzen. Da Sicherungspunkte zusätzlichen Overhead im laufenden Normalbetrieb verursachen, kann der Systemadministrator bei den meisten kommerziellen DBS die Frequenz der Sicherungspunkterzeugung einstellen (typische Werte liegen in der Größenordnung von 10 Minuten).

  Variante 1: Synchrone Sicherungspunkte
  Beim Sicherungspunkt werden alle geänderten Seiten aus dem Datenbankpuffer in die Datenbank auf Platte zurückgeschrieben. Zusätzlich wird die Tatsache, daß ein Sicherungspunkt erzeugt wurde in der Logdatei durch einen Checkpoint-Logsatz vermerkt. Die Plattenadresse (und/oder LSN) des jeweils jüngsten Checkpoint-Logsatzes wird in einem speziellen "Bootstrap-Block" mit einer festen Plattenadresse festgehalten.
  Bei einem eventuellen späteren Warmstart muß die Redo-Phase die Logdatei dann nur ab dem jüngsten Checkpoint-Logsatz lesen.

  Die Analysephase kann ebenfalls optimiert werden, indem man bei einem Sicherungspunkt in den Checkpoint-Logsatz eine Liste der zu diesem Zeitpunkt aktiven Transaktionen aufnimmt sowie für jede aktive Transaktion die LSN des jeweils letzten Logsatzes der Transaktion vor dem Sicherungspunkt. Beim Warmstart braucht die Analysephase dann ebenfalls die Logdatei nur ab dem jüngsten Checkpoint-Logsatz lesen.

  Variante 2: Asynchrone Sicherungspunkte
  Synchrone Sicherungspunkte belasten den laufenden Normalbetrieb vor allem dadurch, daß sie sehr viele Pufferseiten "impulsartig" innerhalb kurzer Zeit in die Datenbank schreiben. Bei asynchrone Sicherungspunkten werden zum Zeitpunkt des Sicherungspunktes gar keine Seiten in die Datenbank zurückgeschrieben. Stattdessen läßt man einen ständig mitlaufenden Hintergrundprozeß veränderte Pufferseiten immer dann in Datenbank zurückschreiben, wenn die entsprechende Platte gerade unbelegt ist.
  Um trotzdem den Analyse- und Redo-Aufwand beim Warmstart gering zu halten, werden bei einem asynchronen Sicherungspunkt Informationen über den aktuellen Pufferinhalt in

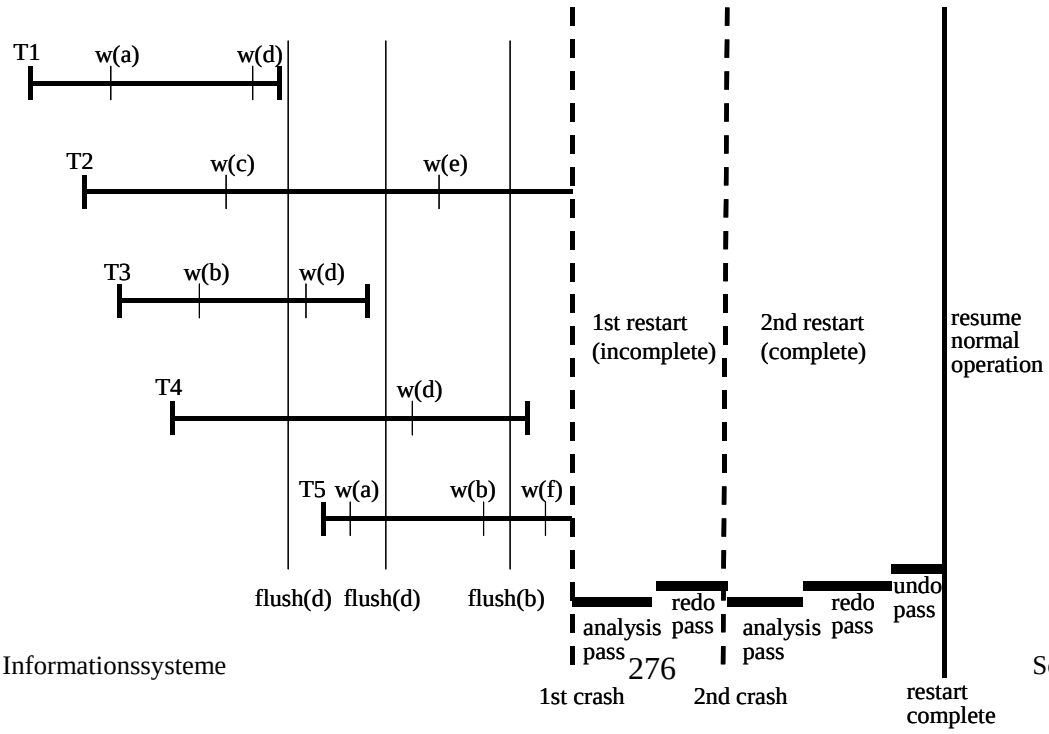
den Checkpoint-Logsatz geschrieben (zusätzlich zur Information über die aktiven Transaktionen): Für jede veränderte Pufferseite wird die LSN des ältesten Logsatzes dieser Seite seit dem letzten Zurückschreiben der Seite im Checkpoint-Logsatz eingetragen.
Beim Warmstart braucht die Redo-Phase die Logdatei dann nur ab der ältesten LSN aller entsprechenden Seiten-Einträge im jüngsten Checkpoint-Logsatz lesen.

Die Analyse-Phase kann außerdem Information über die zum Crash-Zeitpunkt „schmutzigen“ und damit potentiell Redo-bedürftigen Seiten im Cache sowie deren jeweils ältester Redo-bedürftiger Änderung rekonstruieren. Dies Information wird in der sog. Dirty-Pages-Table im Hauptspeicher zusammengestellt; sie zielt auf die Minimierung der Random I/Os auf der Datenbank während der Redo-Phase und erlaubt verschiedene Formen von I/O-Optimierungen (Batching und Prefetching und damit einhergehende Optimierungen der Plattenlatenzzeiten).
Diese Buchführungsinformation ist konservativ, enthält also auch potentiell schmutzige Seiten, die vor dem Crash vom Cache-Manager in die Datenbank zurückgeschrieben wurden. Durch zusätzliches Protokollieren dieser Flush-Aktionen des Cache-Managers kann die DirtyPagesTable-Information noch weiter verbessert werden.

Behandlung subtiler Sonderfälle:

Um das Undo für Transaktionen, die bereits vor dem Crash Verlierer waren korrekt und gleichzeitig möglichst einfach behandeln zu können, werden auch für alle Undo-Schritte selbst Logsätze geschrieben, sog. Compensation Log Entries (CLEs). Beim Redo werden diese Schritte wiederholt ; beim Undo werden sie zusammen mit den Logsätzen der dazugehörigen regulären Vorwärtsoperationen übersprungen.
Diese Redo-History-Methode deckt zum einen den Fall ab, dass Transaktionen im laufenden Betrieb zurückgesetzt wurden (z.B. als Deadlockopfer), bevor es später zu einem Systemabsturz kommt, und zum anderen den Fall, dass das System während der Crash-Recovery selbst nochmals ausfällt. Ohne CLEs wäre die korrekte Behandlung solcher Sonderfälle sehr kompliziert und potentiell ineffizient.

Beispielablauf (ohne Sicherungspunkte):



Sequence number: action	Change of cached database [PageNo: SeqNo]	Change of stable database [PageNo: SeqNo]	Log entry added to log buffer [LogSeqNo: action]	Log entries added to stable log [LogSeqNo's]
1:begin(T1)			1: begin(T1)	
2: begin(T2)			2: begin(T2)	
3: write(a,T1)	a: 3		3: write(a,T1)	
4: begin(T3)			4: begin(T3)	
5: begin(T4)			5: begin(T4)	
6: write(b,T3)	b: 6		6: write(b,T3)	
7: write (c,T2)	c: 7		7: write(c,T2)	
8: write(d,T1)	d: 8		8: write(d,T1)	
9: commit(T1)			9: commit(T1)	1,2,3,4,5,6,7,8,9
10: flush(d)		d:8		
11: write(d,T3)	d: 11		11: write(d,T3)	
12: begin(T5)			12: begin(T5)	
13: write(a,T5)	a: 13		13: write(a,T5)	
14: commit(T3)			14: commit(T3)	11,12,13,14
15: flush(d)		d: 11		
16: write(d,T4)	d: 16		16: write(d,T4)	
17: write(e,T2)	e: 17		17: write(e,T2)	
18: write(b,T5)	b: 18		18: write(b,T5)	
19: flush(b)		b: 18		16,17,18
20: commit(T4)			20: commit(T4)	20
21: write(f,T5)	f: 21		21: write(f,T5)	
system crash				

restart				
analysis pass: losers = {T2,T5}				
redo(3)	a: 3			
consider-redo(6)	b: 18			
flush (a)		a: 3		
consider-redo(8)	d: 11			
consider-redo(11)	d: 11			
second system crash				

second restart				
analysis pass: losers = {T2,T5}				
consider-redo(3)	a:3			
consider-redo(6)	b: 18			
consider-redo(8)	d: 11			
consider-redo(11)	d: 11			
redo(16)	d: 16			
undo(18)	b: 17			
consider-undo(17)	e: 0			
consider-undo(13)	a: 3			
consider-undo(7)	c: 0			
second restart complete: resume normal operation				

Ablauf für dasselbe Beispiel mit Optimierungen (asynchrone Sicherungspunkte, Flush-Logsätze):

Sequence number: action	Change of cached database [PageNo: SeqNo]	Change of stable database [PageNo: SeqNo]	Log entry added to log buffer [LogSeqNo: action]	Log entries added to stable log [LogSeqNo's]
1:begin(T1)			1: begin(T1)	
2: begin(T2)			2: begin(T2)	
3: write(a,T1)	a: 3		3: write(a,T1)	
4: begin(T3)			4: begin(T3)	
5: begin(T4)			5: begin(T4)	
6: write(b,T3)	b: 6		6: write(b,T3)	
7: write (c,T2)	c: 7		7: write(c,T2)	
8: write(d,T1)	d: 8		8: write(d,T1)	
9: commit(T1)			9: commit(T1)	1,2,3,4,5,6,7,8,9
10: flush(d)		d:8	**10: flush(d)**	
11: write(d,T3)	d: 11		11: write(d,T3)	
12: begin(T5)			12: begin(T5)	
13: write(a,T5)	a: 13		13: write(a,T5)	
14: checkpoint			14: CP DirtyPages: {a,b,c,d} RedoLSNs: a:3, b:6, c:7, d:11 ActiveTrans: {T2,T3,T4,T5}	10,11,12,13,14
15: commit(T3)			15: commit(T3)	15
16: flush(d)		d: 11	**16: flush(d)**	
17: write(d,T4)	d: 17		17: write(d,T4)	
18: write(e,T2)	e: 18		18: write(e,T2)	
19: write(b,T5)	b: 19		19: write(b,T5)	
20: flush(b)		b: 19	**20: flush(b)**	16,17,18,19
21: commit(T4)			21: commit(T4)	20,21
22: write(f,T5)	f: 22		22: write(f,T5)	
system crash				

restart				
analysis pass: losers = {T2,T5} **DirtyPages = {a,c,d,e,f}** **RedoLSNs: a:3, c:7, d:17, e:18**				
redo(3)	a:3			
consider-redo(6)	b: 19			
skip-redo(8)				
skip-redo(11)				
redo(17)	d:17			
undo(19)	b: 18			
consider-undo(18)	e: 0			
consider-undo(13)	a: 3			
consider-undo(7)	c: 0			
restart complete: resume normal operation				

Pseudocode für optimierte Crash-Recovery:

```
/* Data structures for the page model recovery algorithm of choice: */

type Page: record of
        PageNo: identifier;
        PageSeqNo: identifier;
        Status: (clean, dirty);
        Contents: array [PageSize] of char;
     end;
persistent var StableDatabase:
        set of Page indexed by PageNo;
var DatabaseCache:
        set of Page indexed by PageNo;
type LogEntry: record of
        LogSeqNo: identifier;
        TransId: identifier;
        PageNo: identifier;
        ActionType: (write, full-write, begin, commit, rollback,
             compensate, checkpoint, flush);
        ActiveTrans: set of TransInfo;
             /* present only in log entries of type checkpoint */
        DirtyPages: set of DirtyPageInfo;
             /* present only in log entries of type checkpoint */
        UndoInfo: array of char;
        RedoInfo: array of char;
        PreviousSeqNo: identifier;
        NextUndoSeqNo: identifier;
     end;
persistent var StableLog:
        ordered set of LogEntry indexed by LogSeqNo;
var LogBuffer:
        ordered set of LogEntry indexed by LogSeqNo;
persistent var MasterRecord: record of
        StartPointer: identifier;
        LastCP: identifier;
     end;
type TransInfo: record of
        TransId: identifier;
        LastSeqNo: identifier;
     end;
var ActiveTrans:
        set of TransInfo indexed by TransId;
typeDirtyPageInfo: record of
        PageNo: identifier;
        RedoSeqNo: identifier;
     end;
var DirtyPages:
        set of DirtyPageInfo indexed by PageNo;
```

```
/* actions during normal operation */

write or full-write (pageno, transid, s):
   DatabaseCache[pageno].Contents := modified contents;
   DatabaseCache[pageno].PageSeqNo := s;
   DatabaseCache[pageno].Status := dirty;
   newlogentry.LogSeqNo := s;
   newlogentry.ActionType := write or full-write;
   newlogentry.TransId := transid;
   newlogentry.PageNo := pageno;
   newlogentry.UndoInfo := information to undo update (before image for full-write);
   newlogentry.RedoInfo := information to redo update (after image for full-write);
   newlogentry.PreviousSeqNo := ActiveTrans[transid].LastSeqNo;
   ActiveTrans[transid].LastSeqNo := s;
   LogBuffer += newlogentry;
   if pageno not in DirtyPages then
      DirtyPages += pageno;
      DirtyPages[pageno].RedoSeqNo := s;
   end /*if*/;

fetch (pageno):
   DatabaseCache += pageno;
   DatabaseCache[pageno].Contents := StableDatabase[pageno].Contents;
   DatabaseCache[pageno].PageSeqNo := StableDatabase[pageno].PageSeqNo;
   DatabaseCache[pageno].Status := clean;

flush (pageno):
   if there is logentry in LogBuffer with logentry.PageNo = pageno
   then
        force ( );
   end /*if*/;
   StableDatabase[pageno].Contents := DatabaseCache[pageno].Contents;
   StableDatabase[pageno].PageSeqNo := DatabaseCache[pageno].PageSeqNo;
   DatabaseCache[pageno].Status := clean;
   newlogentry.LogSeqNo := next sequence number to be generated;
   newlogentry.ActionType := flush;
   newlogentry.PageNo := pageno;
   LogBuffer += newlogentry;
   DirtyPages -= pageno;

force ( ):
     StableLog += LogBuffer;
     LogBuffer := empty;

begin (transid, s):
   ActiveTrans += transid;
   ActiveTrans[transid].LastSeqNo := s;
   newlogentry.LogSeqNo := s;
   newlogentry.ActionType := begin;
   newlogentry.TransId := transid;
   newlogentry.PreviousSeqNo := nil;
   LogBuffer += newlogentry;

commit (transid, s):
   newlogentry.LogSeqNo := s;
   newlogentry.ActionType := commit;
   newlogentry.TransId := transid;
   newlogentry.PreviousSeqNo := ActiveTrans[transid].LastSeqNo;
   LogBuffer += newlogentry;
   ActiveTrans -= transid;
   force ( );
```

```
abort (transid):
   logentry :=
         ActiveTrans[transid].LastSeqNo;
   while logentry is not nil and
         logentry.ActionType = write or full-write
   do
         newlogentry.LogSeqNo := new sequence number;
         newlogentry.ActionType := compensation;
         newlogentry.PreviousSeqNo := ActiveTrans[transid].LastSeqNo;
         newlogentry.RedoInfo := inverse action of the action in logentry;
         newlogentry.NextUndoSeqNo := logentry.PreviousSeqNo;
         ActiveTrans[transid].LastSeqNo := newlogentry.LogSeqNo;
         LogBuffer += newlogentry;
         write (logentry.PageNo) according to logentry.UndoInfo;
         logentry := logentry.PreviousSeqNo;
   end /*while*/
   newlogentry.LogSeqNo := new sequence number;
   newlogentry.ActionType := rollback;
   newlogentry.TransId := transid;
   newlogentry.PreviousSeqNo := ActiveTrans[transid].LastSeqNo;
   newlogentry.NextUndoSeqNo := nil;
   LogBuffer += newlogentry;
   ActiveTrans -= transid;
   force ( );


log truncation ( ):
    OldestUndoLSN :=
         min {i | StableLog[i].TransId is in ActiveTrans};
    SystemRedoLSN := min {DirtyPages[p].RedoSeqNo};
    OldestRedoPage := page p such that
         DirtyPages[p].RedoSeqNo = SystemRedoLSN;
    NewStartPointer := min{OldestUndoLSN, SystemRedoLSN};
    OldStartPointer := MasterRecord.StartPointer;
    while OldStartPointer - NewStartPointer is not sufficiently large
          and SystemRedoLSN < OldestUndoLSN
    do
        flush (OldestRedoPage);
        SystemRedoLSN := min{DatabaseCache[p].RedoLSN};
        OldestRedoPage := page p such that
              DatabaseCache[p].RedoLSN = SystemRedoLSN;
        NewStartPointer := min{OldestUndoLSN, SystemRedoLSN};
    end /*while*/;
    MasterRecord.StartPointer := NewStartPointer;

checkpoint ( ):
    logentry.ActionType := checkpoint;
    logentry.ActiveTrans := ActiveTrans (as maintained in memory);
    logentry.DirtyPages := DirtyPages (as maintained in memory);
    logentry.LogSeqNo := next sequence number to be generated;
    LogBuffer += logentry;
    force ( );
    MasterRecord.LastCP := logentry.LogSeqNo;
```

```
/* recovery algorithms */

restart ( ):
   analysis pass ( ) returns losers, DirtyPages;
   redo pass ( );
   undo pass ( );

analysis pass ( ) returns losers, DirtyPages:
   var losers: set of record
            TransId: identifier;
            LastSeqNo: identifier;
         end indexed by TransId;
   cp := MasterRecord.LastCP;
   losers := StableLog[cp].ActiveTrans;
   DirtyPages := StableLog[cp].DirtyPages;
   max := LogSeqNo of most recent log entry in StableLog;
   for i := cp to max do
        case StableLog[i].ActionType:
            begin:
                  losers += StableLog[i].TransId;
                  losers[StableLog[i].TransId].LastSeqNo := nil;
            commit:
                  losers -= StableLog[i].TransId;
            full-write:
                  losers[StableLog[i].TransId].LastSeqNo := i;
        end /*case*/;

        if StableLog[i].ActionType = write or full-write or compensate
              and StableLog[i].PageNo not in DirtyPages
        then
              DirtyPages += StableLog[i].PageNo;
              DirtyPages[StableLog[i].PageNo].RedoSeqNo := i;
        end /*if*/;
        if StableLog[i].ActionType = flush
        then
               DirtyPages -= StableLog[i].PageNo;
        end /*if*/;
   end /*for*/;

redo pass ( ):
   SystemRedoLSN := min {DirtyPages[p].RedoSeqNo};
   max := LogSeqNo of most recent log entry in StableLog;
   for i := SystemRedoLSN to max do
          if StableLog[i].ActionType = write or full-write or compensate
          then
             pageno = StableLog[i].PageNo;
             if pageno in DirtyPages and
                DirtyPages[pageno].RedoSeqNo < i
             then
                fetch (pageno);
                if DatabaseCache[pageno].PageSeqNo < i
                then
                   read and write (pageno)
                        according to StableLog[i].RedoInfo;
                   DatabaseCache[pageno].PageSeqNo := i;
                end /*if*/;
             end /*if*/;
          end /*if*/;
   end /*for*/;
```

```
undo pass ( ):
   ActiveTrans := empty;
   for each t in losers
   do
         ActiveTrans += t;
         ActiveTrans[t].LastSeqNo := losers[t].LastSeqNo;
   end /*for*/;
   while there exists t in losers
         such that losers[t].LastSeqNo <> nil
   do
        nexttrans := TransNo in losers
             such that losers[nexttrans].LastSeqNo =
             max {losers[x].LastSeqNo | x in losers};
       nextentry := losers[nexttrans].LastSeqNo;

       if StableLog[nextentry].ActionType = compensation
       then
          losers[nexttrans].LastSeqNo := StableLog[nextentry].NextUndoSeqNo;
       end /*if*/;

       if StableLog[nextentry].ActionType = write or full-write
       then
          pageno = StableLog[nextentry].PageNo;
          fetch (pageno);
          if DatabaseCache[pageno].PageSeqNo >= nextentry.LogSeqNo
          then
             newlogentry.LogSeqNo := new sequence number;
             newlogentry.ActionType := compensation;
             newlogentry.PreviousSeqNo := ActiveTrans[transid].LastSeqNo;
             newlogentry.NextUndoSeqNo := nextentry.PreviousSeqNo;
             newlogentry.RedoInfo :=
                  inverse action of the action in nextentry;
             ActiveTrans[transid].LastSeqNo := newlogentry.LogSeqNo;
             LogBuffer += newlogentry;
             read and write (StableLog[nextentry].PageNo)
                  according to StableLog[nextentry].UndoInfo;
             DatabaseCache[pageno].PageSeqNo := newlogentry.LogSeqNo;
         end /*if*/;
         losers[nexttrans].LastSeqNo =
               StableLog[nextentry].PreviousSeqNo;
       end /*if*/;

       if StableLog[nextentry].ActionType = begin
       then
           newlogentry.LogSeqNo := new sequence number;
           newlogentry.ActionType := rollback;
           newlogentry.TransId := StableLog[nextentry].TransId;
           newlogentry.PreviousSeqNo := ActiveTrans[transid].LastSeqNo;
           LogBuffer += newlogentry;
           ActiveTrans -= transid;
           losers -= transid;
       end /*if*/;

   end /*while*/;
   force ( );
```

16.2 Commit-Protokolle für verteilte Transaktionen

Problem der Atomarität in einem verteilten System

Transaktionen können Daten in mehreren Systemen ändern (mehrere Datenbanken desselben DBS oder mehrere DBS). Es ist erheblich schwieriger, die Atomarität einer solchen verteilten Transaktion zu gewährleisten. Zusätzlich zu den Logging-Maßnahmen der beteiligten Systeme ist ein verteiltes Protokoll notwendig, um das Commit oder Rollback der Transaktion in allen Systemen einheitlich durchzuführen. Es darf nicht passieren, dass ein System einen Commit-Logsatz für die Transaktion schreibt, während ein anderes System seinen Teil der Transaktion zurücksetzt, beispielsweise weil dieses System abstürzt. Zusätzlich müssen verschiedene Arten von Nachrichtenfehlern (verlorene Nachrichten, duplizierte Nachrichten) verkraftet werden.

2-Phasen-Commit-Protokoll (2PC)

Eines der beteiligten Systeme wird zum *Koordinator* für den Ablauf des Commit einer Transaktion bestimmt. Es kann z.B. immer das System gewählt werden, von dem aus die Transaktion initiiert wurde. Alle Systeme, in denen die Transaktion Änderungen durchgeführt hat, werden als *Agenten* (oder Teilnehmer) bezeichnet. Das Protokoll läuft dann in zwei Phasen ab:

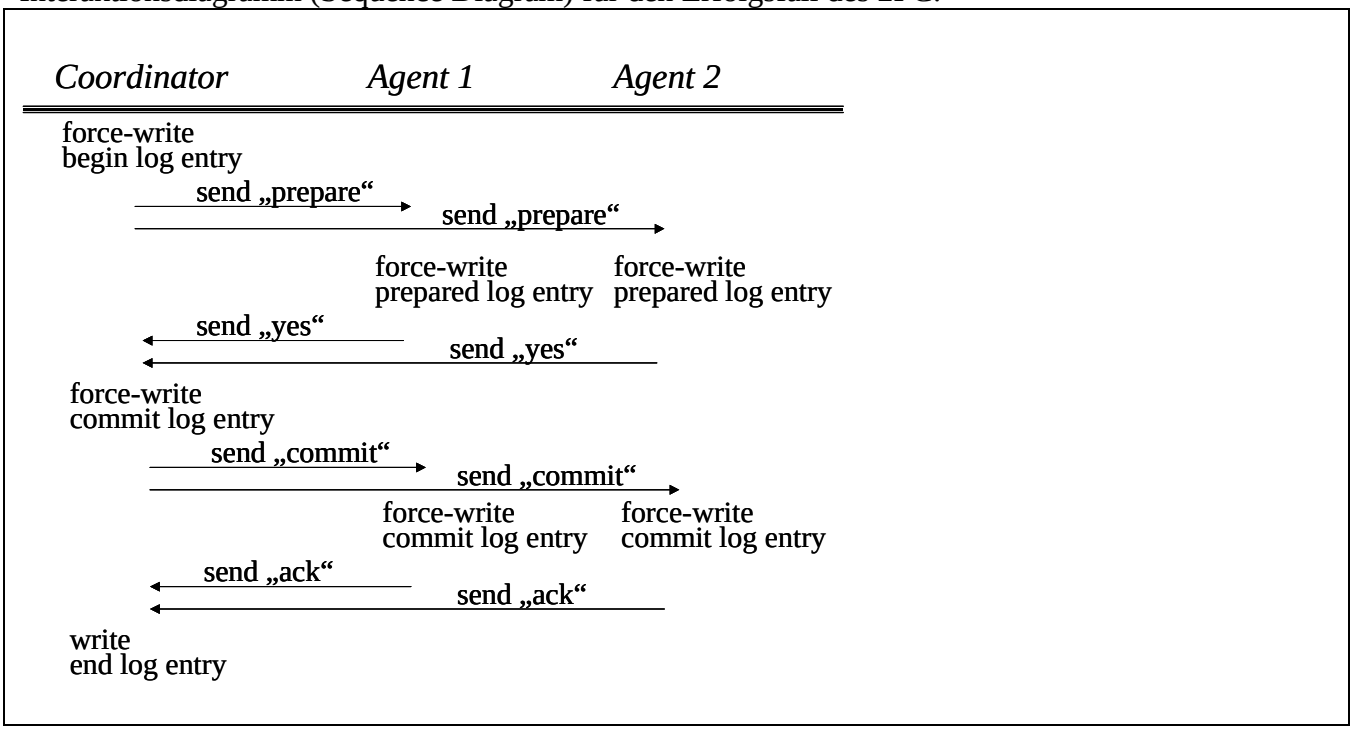
- *Phase 1:* Der Koordinator fragt alle Agenten, ob sie in der Lage sind, die Transaktion mit Commit zu beenden, und bittet die Agenten gleichzeitig, sich sowohl auf ein Commit als auch ein Rollback vorzubereiten („Prepare“-Nachricht). Im positiven Fall gehen dann alle Agenten in den „Prepared“-Zustand über.
- *Phase 2:* Aufgrund des „Umfrageergebnisses“ der ersten Phase trifft der Koordinator dann eine globale Entscheidung für die Transaktion und teilt diese allen Agenten mit. Nur wenn alle Agenten positiv geantwortet haben, kann der Koordinator eine Commit-Entscheidung treffen, andernfalls muss seine Entscheidung Rollback lauten.

In der Zeit zwischen dem Antworten eines Agenten und dem Erhalt der Koordinatorentscheidung kann ein Agent keine eigenmächtige Entscheidung treffen. Insbesondere muss der Agent weiterhin alle Sperren für die Transaktion halten, da auch ein Rollback noch möglich ist. Da der Koordinator ausfallen kann oder die Nachricht mit seiner Entscheidung verloren gehen kann, können Agenten auf diese Weise in eine gewisse Blockierungssituation geraten, die sich sehr negativ auf die Leistung des betroffenen Systems auswirken kann. Man bezeichnet aus diesem Grund das 2PC auch als ein „blockierendes“ Commit-Protokoll. Man kann beweisen, dass es – unter Berücksichtigung aller möglichen Fehlerfälle – kein nichtblockierendes verteiltes Commit-Protokoll geben kann.

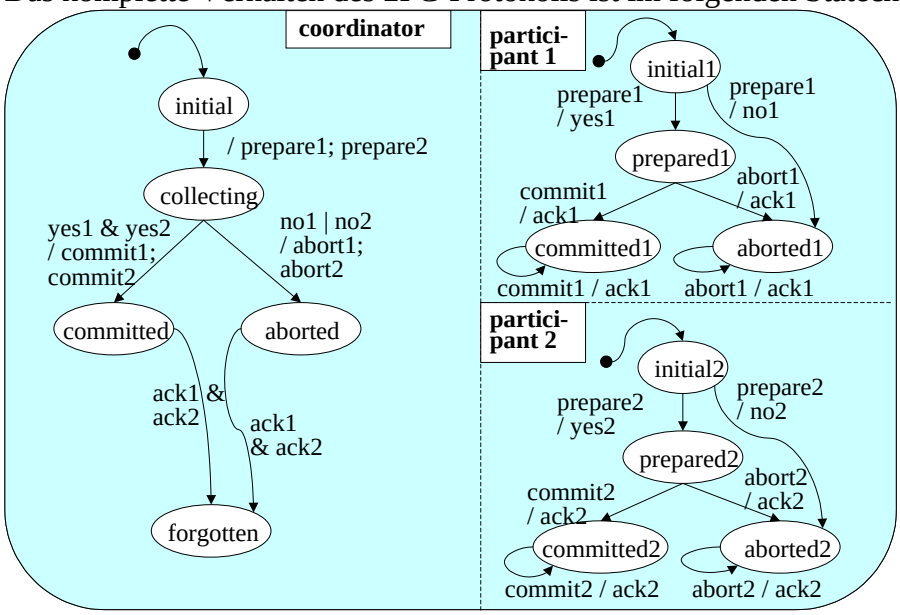
Zusatzbemerkungen:

- Es gibt optimierte Varianten des 2PC, die Nachrichten und/oder Log-I/Os in bestimmten Fällen vermeiden.
- Das 2PC ist standardisiert. Es wird von praktisch allen kommerziellen DBS unterstützt sowie von vielen TP-Monitoren (Transaction Processing Monitors) und Object-Request-Brokers (ORBs) unterstützt. TP-Monitor und ORBs sind Systemsoftware-Komponenten, die eine transaktionsorientierte Kommunikation zwischen Clients und Servers steuern; sie übernehmen dabei typischerweise die Koordinatorrolle im 2PC. ORBs koordinieren Aufrufe von Methoden auf Business-Objekten, die z.B. als sog. Enterprise Java Beans (EJB) gekapselt werden; die Unterstützung von 2PC ist dabei zur Konsistenzerhaltung essentiell.

Interaktionsdiagramm (Sequence Diagram) für den Erfolgsfall des 2PC:



Das komplette Verhalten des 2PC-Protokolls ist im folgenden Statechart spezifiziert (vgl. Kapitel 13):



Behandlung von Nachrichtenverlusten:

Falls bei einem Prozess eine – aufgrund des Protokolls – erwartete Nachricht nicht innerhalb einer bestimmten „Timeout“-Frist eintrifft, wird angenommen, dass der Sender ausgefallen ist, und der Empfänger verhält sich wie einem Knotenausfall.

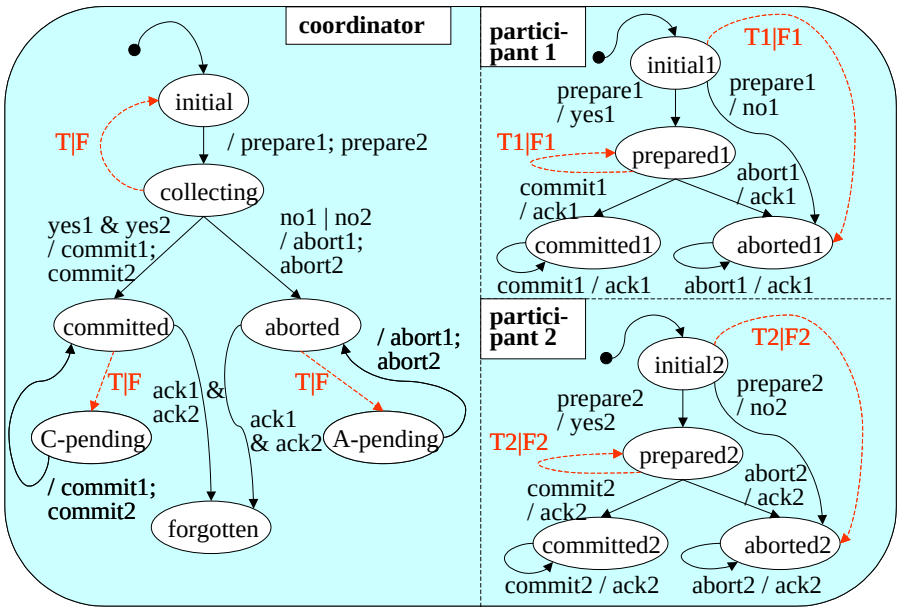
Behandlung von Knotenausfällen:

Ausfallender Prozess	*Reaktion Koordinator*	*Reaktion Agent*
Koordinator in Phase 1 (vor dem Schreiben des Commit-Logsatzes)	Nach dem Restart: Wiederholung des Sendens der Prepare-Nachricht	- Falls noch nicht prepared: Warten oder einseitiges Abort - Falls bereits prepared: Warten (Blockierung) oder ggf. andere Agenten fragen
Koordinator in Phase 2 (vor dem Schreiben des End-Logsatzes)	Nach dem Restart: Wiederholen des Sendens der Commit- oder Abort-Nachricht	- Vor dem Schreiben des Commit- oder Rollback-Logsatzes: Warten (Blockierung) oder ggf. andere Agenten fragen - Nach dem Schreiben des Commit- oder Rollback-Logsatzes: Wiederholen der Ack-Nachricht bei wiederholtem Empfang der Commit- oder Abort-Nachricht
Agent in Phase 1 (vor dem Schreiben des Prepared-Logsatzes)	Bei ausbleibendem Votum: Wiederholen des Sendens der Prepare-Nachricht	Nach dem Restart: Einseitiges Abort
Agent in Phase 2 (vor dem Schreiben des Commit- oder Rollback-Logsatzes)	Bei ausbleibendem Ack: Wiederholen des Sendens der Commit- oder Abort-Nachricht	Nach dem Restart: Warten auf Nachricht des Koordinators (wiederholte Prepare-Nachricht oder (wiederholte) Commit- oder Abort-Nachricht) oder Nachfragen beim Koordinator oder anderen Agenten

Timeouts und eigene Ausfälle können durch ein sog. *Terminationsprotokoll* als Erweiterung des 2PC systematisch in das Protokoll eingebaut werden. Dazu führt man im Statechart zwei zusätzliche Typen von Transitionen ein:

- *Timeout-Transitionen (T)* feuern, wenn ein Prozess zu lange auf eine Nachricht wartet, und
- *Failure-Transitionen (F)* feuern, wenn ein Prozess nach einem Absturz einen Warmstart durchführt und den jüngsten, durch einen Logsatz dokumentierten Zustand vor dem Crash wiederherstellt.

Das vollständige 2PC mitsamt Terminationsprotokoll ist im folgenden Statechart spezifiziert:



Weiterführende Literatur zu Kapitel 16:

- G. Weikum, G. Vossen: Transactional Information Systems – Theory, Algorithms, and the Practice of Concurrency Control and Recovery, Morgan Kaufmann, 2001
- J. Gray, A. Reuter: Transaction Processing – Concepts and Techniques, Morgan Kaufman, 1993