



FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Datenbanksysteme II

Prof. Dr. Viktor Leis

Professur für Datenbanken und Informationssysteme

Verteilte Datenbanken

- Eine *verteilte Datenbank* (VDBMS) ist eine Sammlung von Informationseinheiten die auf verschiedene Rechner verteilt ist, die durch Kommunikationsnetze verbunden sind
- Jede Station kann
 - autonom mit lokalen Daten arbeiten
 - global mit anderen Rechnern des Netzes zusammenarbeiten

Warum Verteilte Datenbanken?

- Daten zu groß für einen Rechner
- Leistung eines Rechners reicht nicht aus
- Mehrere kleine Rechner können günstiger als ein großer Rechner sein
- Ausfallsicherheit

- Shared-Disk: die Festplatte wird von mehreren Rechnern geteilt
- Shared-Memory: alle Rechner teilen sich den Arbeitsspeicher
- Shared-Nothing: unabhängige Rechner, die per Netzwerk kommunizieren
- Shared-Nothing ist die Standardarchitektur (wobei jeder Rechner heute mehrere CPU Kerne hat, die sich den Arbeitsspeicher teilen)

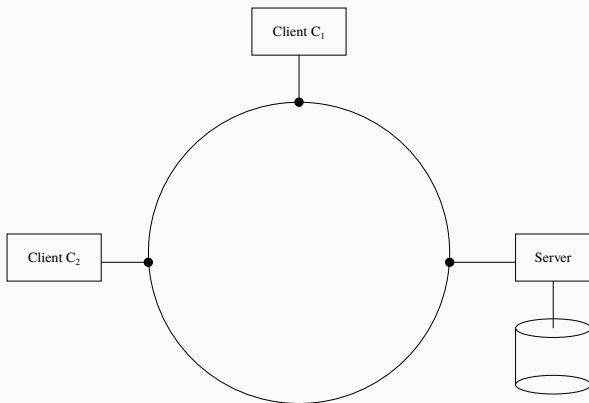
- eine verteilte Datenbank kann nur 2 der folgenden 3 Eigenschaften garantieren:
 - Consistency: Lesezugriffe sehen immer die aktuelle Version
 - Availability: Lese- und Schreibzugriffe sind immer erfolgreich
 - Partition-Tolerance: Das System funktioniert weiter nachdem ein oder mehrere Rechner vom Rest getrennt werden

- Bei dem Kommunikationsnetz kann es sich handeln um
 - LAN: local area network (Ethernet)
 - WAN: wide area network (Internet)
- Kommunikationsnetz ist transparent für Datenbankanwendung

- Durchsatz
 - typisch sind 1 Gbit/s (ca. 100 MB/s)
 - 10 Gbit/s (ca. 1 GB/s) setzt sich nur langsam durch
- Latenz (ping, gemessen vom Uni Jena Netz):
 - Rechner im Nachbarzimmer (selber Switch): 0.1 ms
 - uni-jena.de: 0.6 ms
 - uni-leipzig.de: 2.0 ms
 - tum.de: 10 ms
 - columbia.edu: 107 ms
 - berkeley.edu: 180 ms

Abgrenzung: Client-Server-Architektur

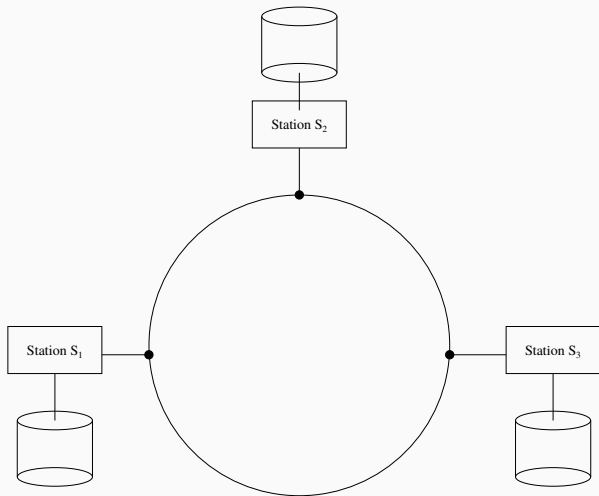
- VDBMS ist keine Client-Server-Architektur



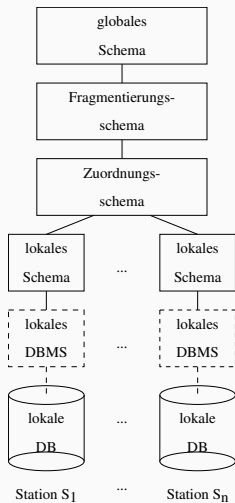
Abgrenzung: Log Shipping

- Schreibzugriffe passieren immer auf dem primären System
- Änderungen werden (synchron oder asynchron) auf weitere sekundäre Datenbankinstanz(en) repliziert
- Replikation passiert über DBMS-log (“Log Shipping”)
- Alle beteiligten Rechner haben eine komplette Kopie der Datenbank
- Bei einem Ausfall des primären Rechners kann auf einen sekundären umgeschaltet werden (üblicherweise passiert dies manuell)
- Lesezugriffe können auch auf sekundären Rechnern ausgeführt werden

- Jede Station hält eigene Daten

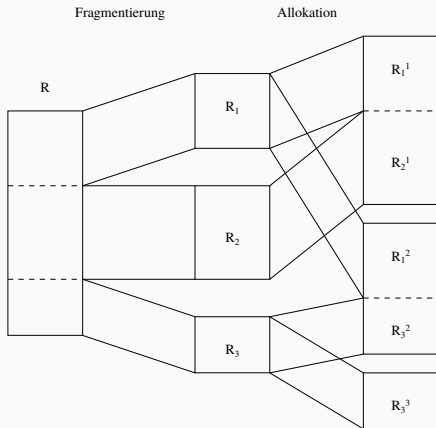


Aufbau eines VDBMS



- Fragmentierung
 - Fragmente enthalten Daten mit gleichem Zugriffsverhalten
- Allokation
 - Fragmente werden den Stationen zugeordnet
 - Mit Replikation
 - Ohne Replikation

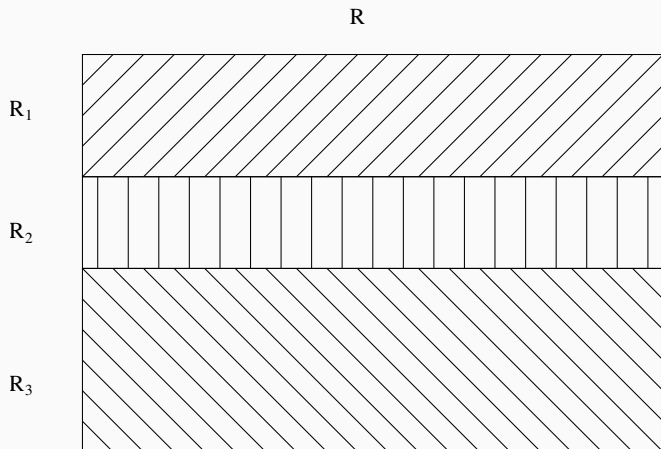
Fragmentierung/Allokation (2)



- Es existieren verschiedene Methoden der Fragmentierung:
 - Horizontal: Zerlegung einer Relation in disjunkte Tupelmengen, Zerlegung durch Selektionen
 - Vertikal: Zusammenfassen von Attributen mit gleichen Zugriffsmustern, Zerlegung durch Projektionen
 - Kombiniert: horizontale und vertikale Fragmentierung auf der gleichen Relation

- Es gibt drei grundlegende Korrektheitsanforderungen an Fragmentierungen:
 - Rekonstruierbarkeit: die Ursprungsrelation läßt sich aus den Fragmenten wiederherstellen
 - Vollständigkeit: jedes Datum ist einem Fragment zugeordnet
 - Disjunktheit: Fragmente überlappen sich nicht, d.h. ein Datum ist nicht mehreren Fragmenten zugeordnet

Horizontale Fragmentierung



Horizontale Fragmentierung (2)

- Bei n Zerlegungsprädikaten gibt es insgesamt 2^n mögliche Fragmente
- Ein Prädikat p_1 :

$$R_1 := \sigma_{p_1}(R)$$

$$R_2 := \sigma_{\neg p_1}(R)$$

- Zwei Prädikate p_1, p_2 :

$$R_1 := \sigma_{p_1 \wedge p_2}(R)$$

$$R_2 := \sigma_{p_1 \wedge \neg p_2}(R)$$

$$R_3 := \sigma_{\neg p_1 \wedge p_2}(R)$$

$$R_4 := \sigma_{\neg p_1 \wedge \neg p_2}(R)$$

Professoren						
PersNr	Name	Rang	Raum	Fakultät	Gehalt	Steuerklasse
2125	Sokrates	C4	226	Philosophie	85000	1
2126	Russel	C4	232	Philosophie	80000	3
2127	Kopernikus	C3	310	Physik	65000	5
2133	Popper	C3	52	Philosophie	68000	1
2134	Augustinus	C3	309	Theologie	55000	5
2136	Curie	C4	36	Physik	95000	3
2137	Kant	C4	7	Philosophie	98000	1

Beispiel (2)

$\rho_1 \equiv \text{Fakultät} = \text{'Theologie'}$

$\rho_2 \equiv \text{Fakultät} = \text{'Physik'}$

$\rho_3 \equiv \text{Fakultät} = \text{'Philosophie'}$

$\text{TheolProfs}' := \sigma_{\rho_1 \wedge \neg \rho_2 \wedge \neg \rho_3}(\text{Professoren}) = \sigma_{\rho_1}(\text{Professoren})$

$\text{PhysikProfs}' := \sigma_{\neg \rho_1 \wedge \rho_2 \wedge \neg \rho_3}(\text{Professoren}) = \sigma_{\rho_2}(\text{Professoren})$

$\text{PhiloProfs}' := \sigma_{\neg \rho_1 \wedge \neg \rho_2 \wedge \rho_3}(\text{Professoren}) = \sigma_{\rho_3}(\text{Professoren})$

$\text{AndereProfs}' := \sigma_{\neg \rho_1 \wedge \neg \rho_2 \wedge \neg \rho_3}(\text{Professoren})$

Abgeleitete horizontale Fragmentierung

- Manchmal ist es sinnvoll eine Relation abhängig von einer anderen horizontalen Fragmentierung zu zerlegen
- Beispiel: völlig unabhängige Zerlegung von Vorlesungen nach SWS:

$$2\text{SWSVorls} := \sigma_{\text{SWS}=2}(\text{Vorlesungen})$$

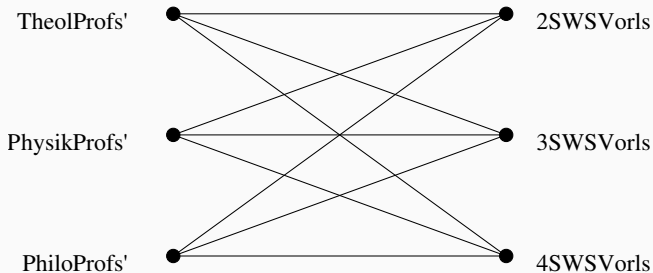
$$3\text{SWSVorls} := \sigma_{\text{SWS}=3}(\text{Vorlesungen})$$

$$4\text{SWSVorls} := \sigma_{\text{SWS}=4}(\text{Vorlesungen})$$

Abgeleitete Fragmentierung (2)

- Bei Beantwortung folgender Anfrage müssen 9 Joins von Fragmenten durchgeführt werden:

```
select Title, Name  
from  Vorlesungen, Professoren  
where gelesenVon=PersNr
```



Abgeleitete Fragmentierung (3)

- Sinnvoller ist folgende (abgeleitete) Fragmentierung:

TheoVorls := Vorlesungen $\bowtie_{\text{gelesenVon=PersNr}}$ TheoProf's'

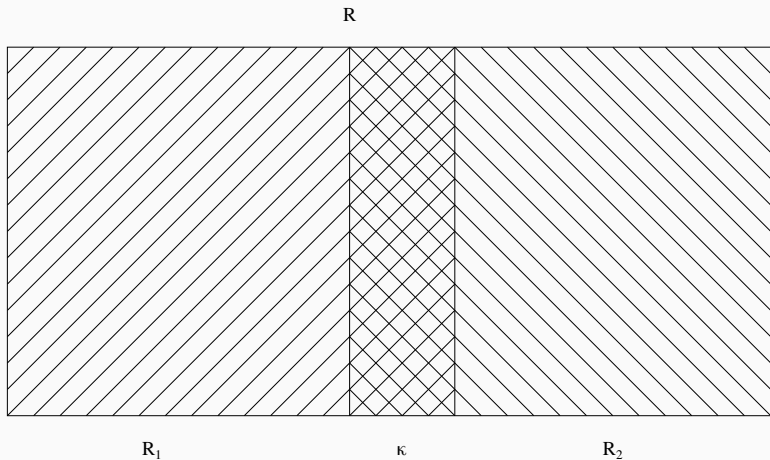
PhysikVorls := Vorlesungen $\bowtie_{\text{gelesenVon=PersNr}}$ PhysikProf's'

PhiloVorls := Vorlesungen $\bowtie_{\text{gelesenVon=PersNr}}$ PhiloProf's'

TheoProf's' ●—————● TheoVorls

PhysikProf's' ●—————● PhysikVorls

PhiloProf's' ●—————● PhiloVorls



- Bei Zerlegung ohne Überlappung gibt es bei vertikaler Fragmentierung ein Problem: Verstoß gegen die Rekonstruierbarkeit
- Man lässt "leichten" Verstoß gegen Disjunktheit zu:
 - Jedes Fragment enthält Primärschlüssel
 - Jedem Tupel der Originalrelation wird künstlicher Surrogatschlüssel zugewiesen, der in Fragment übernommen wird

- Ein Fragment für die Univerwaltung: ProfVerw
- Ein Fragment für Lehre und Forschung: Profs

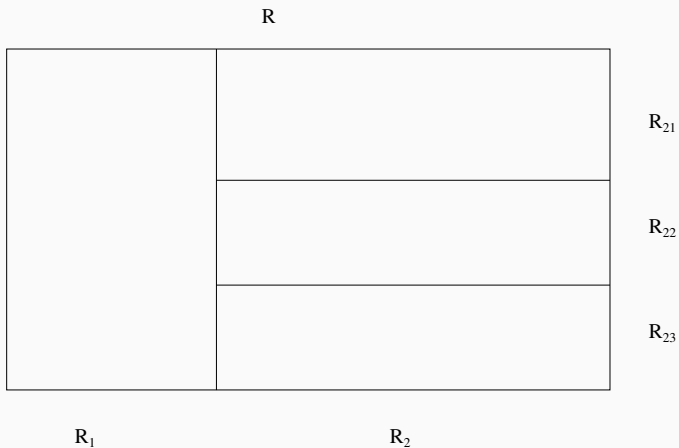
$\text{ProfVerw} := \Pi_{\text{PersNr, Name, Gehalt, Steuerklasse}}(\text{Professoren})$

$\text{Profs} := \Pi_{\text{PersNr, Name, Rang, Raum, Fakultät}}(\text{Professoren})$

$\text{Professoren} = \text{ProfVerw} \bowtie_{\text{ProfVerw.PersNr}=\text{Profs.PersNr}} \text{Profs}$

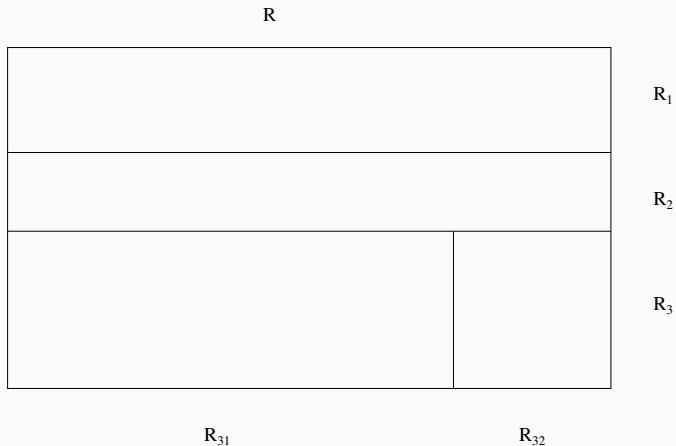
Kombinierte Fragmentierung

- Erst vertikal, dann horizontal:

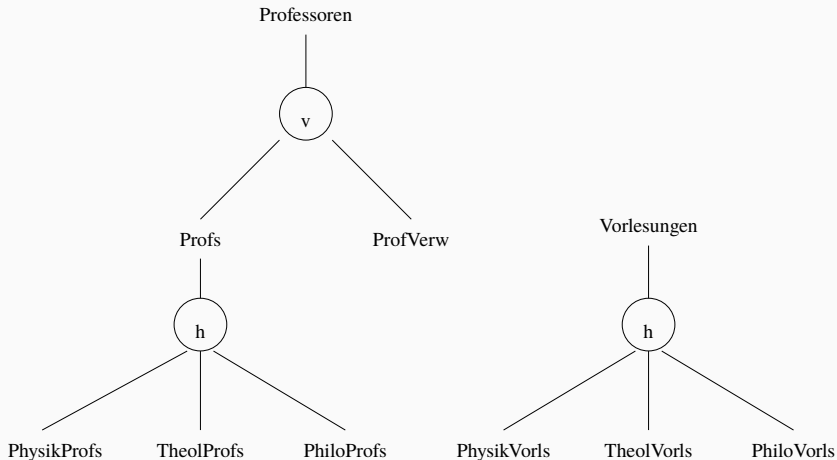


Kombin. Fragmentierung (2)

- Erst horizontal, dann vertikal:



Beispielanwendung



Beispielanwendung (2)

- Bei der Allokation werden nun die Fragmente Stationen zugeteilt (hier ohne Replikation)

Station	Bemerkung	zugeordnete Fragmente
S_{Verw}	Verwaltungsrechner	$\{ProfVerw\}$
S_{Physik}	Dekanat Physik	$\{PhysikVorls, PhysikProfs\}$
S_{Philo}	Dekanat Philosophie	$\{PhiloVorls, PhiloProfs\}$
S_{Theol}	Dekanat Theologie	$\{TheolVorls, TheolProfs\}$

- Unter *Transparenz* versteht man den Grad an Unabhängigkeit, den ein VDBMS dem Benutzer vermittelt
- Es werden verschiedene Stufen unterschieden:
 - Fragmentierungstransparenz
 - Allokationstransparenz
 - Lokale Schema-Transparenz

- Höchste Stufe der Transparenz (Idealzustand)
- Benutzer arbeitet auf globalem Schema und VDBMS übersetzt Anfragen in Operationen auf Fragmenten
- Beispiel:

```
select Titel, Name  
from Vorlesungen, Professoren  
where gelesenVon = PersNr
```


- Nächst niedrigere Stufe
- Benutzer muss zwar Fragmente kennen, aber nicht deren Aufenthaltsort
- Beispiel:

```
select Gehalt  
from ProfVerw  
where Name = 'Sokrates'
```

- Bei dieser Stufe muss Benutzer sowohl Fragment also auch Aufenthaltsort kennen
- Es stellt sich die Frage, inwieweit überhaupt noch Transparenz vorliegt (alle Rechner benutzen das selbe Datenmodell)
- Beispiel:

```
select Name  
from   TheolProfs at  $S_{Theol}$   
where Rang = 'C3'
```

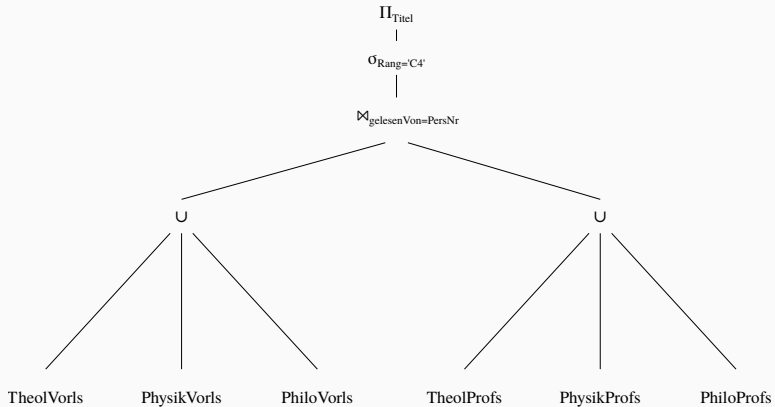
- Durch die Verteilung der Daten müssen folgende Bereiche angepasst werden:
 - Anfragebearbeitung/-optimierung
 - Transaktionskontrolle
 - Mehrbenutzersynchronisation

- Wir unterscheiden zwischen
 - Horizontaler Fragmentierung
 - Vertikaler Fragmentierung

```
select Titel
from Vorlesungen, Profs
where gelesenVon = PersNr and
      Rang = 'C3'
```

- Rekonstruiere alle in der Anfrage vorkommenden globalen Relationen aus den Fragmenten
- Kombiniere den Rekonstruktionsausdruck mit dem Ausdruck aus der Übersetzung der SQL-Anfrage

Kanonische Form



- Kanonische Form ist zwar korrekt, aber ineffizient
- Eine zentrale Eigenschaft der relationalen Algebra ist:

$$(R_1 \cup R_2) \bowtie_p (S_1 \cup S_2) = (R_1 \bowtie_p S_1) \cup (R_1 \bowtie_p S_2) \cup (R_2 \bowtie_p S_1) \cup (R_2 \bowtie_p S_2)$$

- Damit ist aber nicht viel erreicht (für das Zusammensetzen von R_1, \dots, R_n und S_1, \dots, S_m sind $n \cdot m$ Joinoperationen nötig)

Optimierung (2)

- Wenn aber jedes S_i eine abgeleitete horizontale Fragmentierung ist, d.h.

$$S_i = S \bowtie_p R_i \text{ mit } S = S_1 \cup \dots \cup S_n$$

- dann gilt

$$R_i \bowtie_p S_j = \emptyset \text{ für } i \neq j$$

- und somit

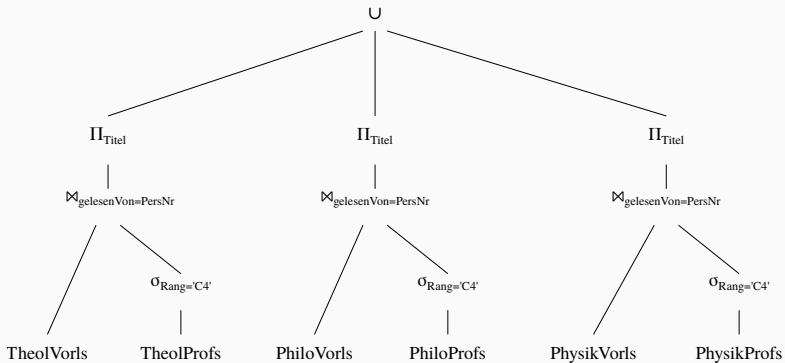
$$(R_1 \cup \dots \cup R_n) \bowtie_p (S_1 \cup \dots \cup S_n) = (R_1 \bowtie_p S_1) \cup (R_2 \bowtie_p S_2) \cup \dots \cup (R_n \bowtie_p S_n)$$

- Damit können die Joins aus unserem Beispiel lokal ausgeführt werden
- Außerdem gibt es Regeln, um Selektionen und Projektionen nach unten zu schieben:

$$\sigma_p(R_1 \cup R_2) = \sigma_p(R_1) \cup \sigma_p(R_2)$$

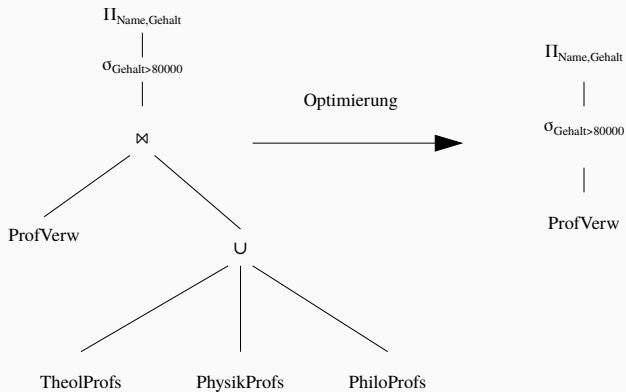
$$\Pi_L(R_1 \cup R_2) = \Pi_L(R_1) \cup \Pi_L(R_2)$$

Optimierter Plan



```
select Name, Gehalt  
from Professoren  
where Gehalt>80000
```

- Naiver Ansatz: globale Relation rekonstruieren, dann Anfrage auswerten
- Sinnvoller: nur relevante Fragmente holen



- Problem: zu joinende Relationen können auf verschiedenen Stationen liegen
- Betrachtung des allgemeinsten Falls:
 - Äußere Relation R ist auf Station St_R
 - Innere Relation S ist auf Station St_S
 - Ergebnis wird auf Station St_{Result} benötigt

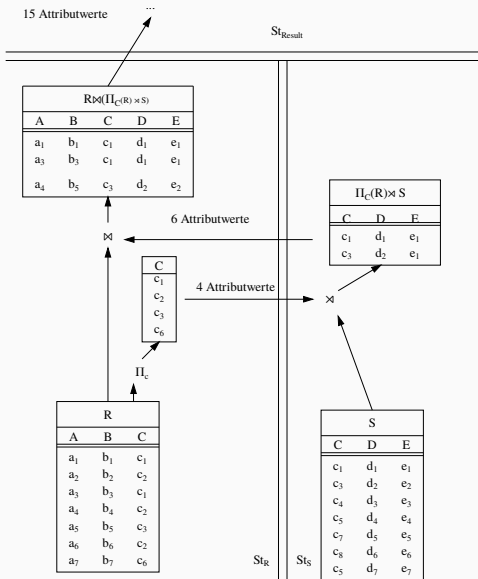
- Nested Loop: iteriere durch Tupel von R , schicke jedes Tupel zu St_S , suche passende Tupel, joine und schicke Ergebnis nach St_{Result}
- Transfer einer Relation: schicke komplette Relation zum anderen Knoten und führe dort Join aus, schicke Ergebnis nach St_{Result}
- Transfer beider Relationen: schicke beide Relationen zu St_{Result} und führe dort den Join aus

- Ohne Filterung müssen große Datenmengen über das Netz, obwohl Ergebnis eventuell sehr klein
- Idee: verschicke nur Tupel, die auch Joinpartner finden
- Folgende Eigenschaften werden dabei genutzt (C ist Joinattribut):

$$R \bowtie S = R \bowtie (R \bowtie S)$$

$$R \bowtie S = \Pi_C(R) \bowtie S$$

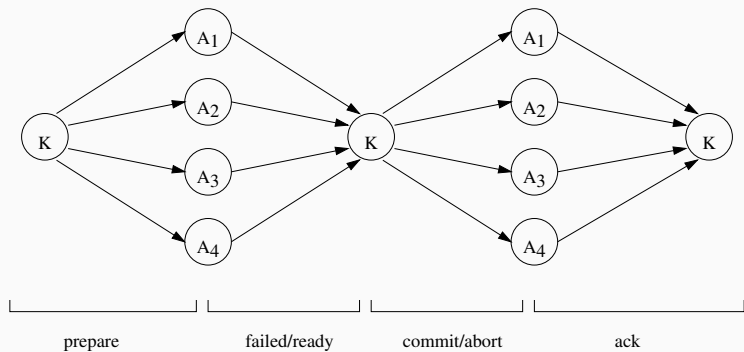
Beispiel



- Transaktionen können sich über mehrere Rechnerknoten erstrecken
- Alle Stationen schreiben lokale Protokolleinträge über ausgeführte Operationen
- Wird beim Wiederanlauf benötigt, um Daten einer abgestürzten Station zu rekonstruieren

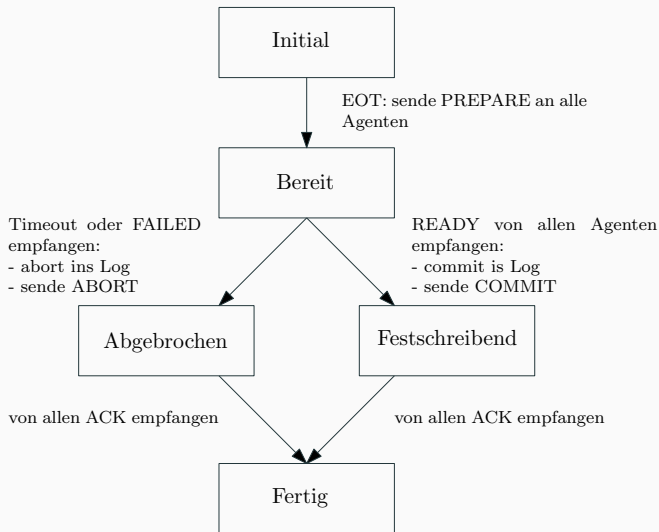
- Abort: bei einem Abbruch einer globalen Transaktion müssen alle lokalen Teile zurückgesetzt werden
- Commit: prinzipielle Schwierigkeit beim Beenden:
 - Atomare Beendigung der Transaktion muss gewährleistet sein

Two-Phase Commit (2PC)

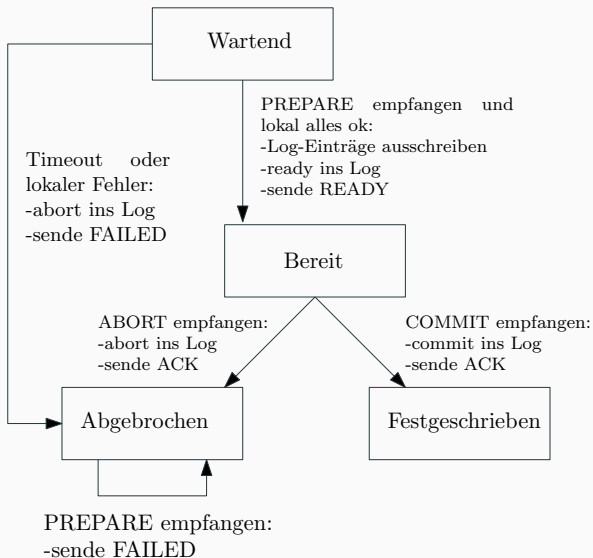


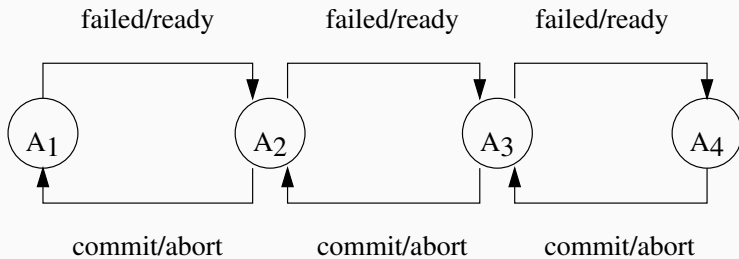
K=Koordinator, A=Agent

Zustandsübergänge Koordinator



Zustandsübergänge Agent





- Während des verteilten Commits kann es zu folgenden Fehlerfällen kommen:
 - Absturz des Koordinators
 - Absturz eines Agenten
 - Verlorene Nachricht

Verteilte Mehrbenutzersynchronisation

- Lokale Serialisierbarkeit garantiert noch keine globale Serialisierbarkeit
- Beispiel:

S_1

Schritt	T_1	T_2
1.	$r(A)$	
2.		$w(A)$

S_2

Schritt	T_1	T_2
3.		$w(B)$
4.	$r(B)$	

- Reguläres 2PL reicht auch im verteilten Fall nicht aus
- Erst strenges 2PL garantiert Serialisierbarkeit
- Verwaltung der Sperren:
 - zentral
 - lokal

- Alle Transaktionen fordern Sperren auf einer dedizierten Station an
- Diese Station kann leicht zum “Bottleneck” werden
- Außerdem verstößt dieses Verfahren gegen lokale Autonomie der Stationen
- Deswegen wird dieses Verfahren nicht angewendet

- Globale Transaktionen (TAs die auf mehr als einer Station laufen) müssen sich vor Modifikation eines Datenelements die Sperre vom lokalen Sperrverwalter holen
- Lokale Transaktionen müssen nur mit ihrem eigenen Verwalter kommunizieren
- Erkennung von Deadlocks ist allerdings schwieriger als bei der zentralen Verwaltung

- Eine lokale Deadlockerkennung reicht nicht:

S ₁		
Schritt	T ₁	T ₂
0.	BOT	
1.	lockS(A)	
2.	r(A)	
6.		lockX(A) ~~~~~

S ₂		
Schritt	T ₁	T ₂
3.		BOT
4.		lockX(B)
5.		w(B)
7.	lockS(B) ~~~~~	

- Timeouts: nach Verstreichen eines Zeitintervalls wird TA zurückgesetzt (Wahl des Intervalls kritisch)
- Zentralisierte Deadlockerkennung: ein Knoten baut einen zentralen Wartegraphen (hoher Aufwand, Phantomdeadlocks)
- Dezentrale Deadlockerkennung: Lokale Wartegraphen + spezieller Knoten *External*

- Jeder TA wird ein Heimatknoten zugeordnet (üblicherweise dort wo TA begonnen wurde)
- Eine TA kann externe Subtransaktionen auf anderen Stationen starten
- In Deadlockbeispiel ist S_1 Heimat von T_1 und S_2 Heimat von T_2

Dezentrale Erkennung (2)

- Für eine externe Subtransaktion T_i wird folgende Kante eingeführt:

$$External \rightarrow T_i$$

- Auf einer anderen Station wird auf Fertigstellung von T_i gewartet (nämlich von der TA, die die externe Subtransaktion initiiert hat)
- Für eine TA T_j die eine Subtransaktion initiiert die Kante

$$T_j \rightarrow External$$

- T_j wartet auf Fertigstellung der auf einer anderen Station angestoßenen Subtransaktion

Dezentrale Erkennung (3)

- Für unser Beispiel bedeutet dies

$$S_1 : \boxed{External \rightarrow T_2 \rightarrow T_1 \rightarrow External}$$

$$S_2 : \boxed{External \rightarrow T_1 \rightarrow T_2 \rightarrow External}$$

- Ein Zyklus der *External* enthält ist nicht notwendigerweise ein Deadlock
- Zur Feststellung eines Deadlocks müssen Stationen Informationen austauschen

Dezentrale Erkennung (4)

- Station mit lokalem Wartegraph

$$External \rightarrow T'_1 \rightarrow T'_2 \rightarrow \dots \rightarrow T'_n \rightarrow External$$

schickt ihren lokalen Graphen an die Station, wo T'_n eine Subtransaktion angestoßen hat

- Für unser Beispiel:

$$S_2 : \boxed{External \begin{matrix} \rightarrow \\ \leftarrow \end{matrix} T_1 \begin{matrix} \rightarrow \\ \leftarrow \end{matrix} T_2 \begin{matrix} \rightarrow \\ \leftarrow \end{matrix} External}$$

$$T_1 \rightarrow T_2 \rightarrow T_1$$

$$T_2 \rightarrow T_1 \rightarrow T_2$$

Dezentrale Erkennung (5)

- Um redundante Nachrichten zu vermeiden (im obigen Beispiel schickt S_1 Informationen an S_2 und umgekehrt), wird nicht immer der gesamte Graph verschickt
- Bei (lokalem) Wartegraph

$$External \rightarrow T'_1 \rightarrow T'_2 \rightarrow \dots \rightarrow T'_n \rightarrow External$$

wird Information nur verschickt, wenn TA-Identifikator von T'_n größer ist als TA-Identifikator von T'_1

- Es gibt Verfahren, die Zeitstempel einsetzen, um Deadlocks zu vermeiden
 - Zeitstempelbasierte Synchronisation
 - Deadlockvermeidung bei sperrbasierten Verfahren:
wound/wait, wait/die
- Setzt voraus, daß global eindeutige Zeitstempel generiert werden können

- Gängigste Methode:



- Die Stations-ID muss in den niedrigwertigsten Bits stehen
- Ansonsten würden immer TAs bestimmter Stationen bevorzugt
- Außerdem sollten Uhren nicht zu weit voneinander abweichen

- Was ist, wenn es mehrere Kopien eines Datenelements gibt?
- Wenn immer nur gelesen wird, ist dies unproblematisch
- Es reicht irgendeine Kopie zu lesen
- Problematisch wird es bei Änderungen

- Bei einer Änderungsoperation müssen alle Kopien angepasst werden
- Favorisiert Leseoperationen, hier muss nur eine Kopie gelesen werden
- Bei Ausfall einer Kopie können Änderungsoperationen nicht mehr ausgeführt werden bzw. werden verzögert

- Idee: Kopien bekommen Gewichte (je nach Robustheit und Leistung der Station)
- Es reicht, Kopien mit einem bestimmten Gesamtgewicht einzusammeln

Station (S_i)	Kopie (A_i)	Gewicht (w_i)
S_1	A_1	3
S_2	A_2	1
S_3	A_3	2
S_4	A_4	2

Quorum-Concensus (2)

$$W(A) = \sum_{i=1}^4 w_i(A) = 8.$$

Lesequorum $Q_r(A)$

Schreibquorum $Q_w(A)$

- $Q_w(A) + Q_w(A) > W(A)$ und
- $Q_r(A) + Q_w(A) > W(A)$.

Beispiel:

- $Q_r(A) = 4$
- $Q_w(A) = 5$

Änderungsoperation

- Vor dem Schreiben:

Station	Kopie	Gewicht	Wert	Versions#
S_1	A_1	3	1000	1
S_2	A_2	1	1000	1
S_3	A_3	2	1000	1
S_4	A_4	2	1000	1

- Nach dem Schreiben:

Station	Kopie	Gewicht	Wert	Versions#
S_1	A_1	3	1100	2
S_2	A_2	1	1000	1
S_3	A_3	2	1100	2
S_4	A_4	2	1000	1

- In verteilten Datenbanksystemen werden die Daten auf räumlich (weit) getrennte Rechner verteilt
- Durch die Verteilung der Daten werden einige der üblich verwendeten Mechanismen in DBMS wesentlich komplizierter
- Das Netzwerk verhindert oft gute Geschwindigkeit (sowohl Latenz als auch Bandbreite sind eingeschränkt)