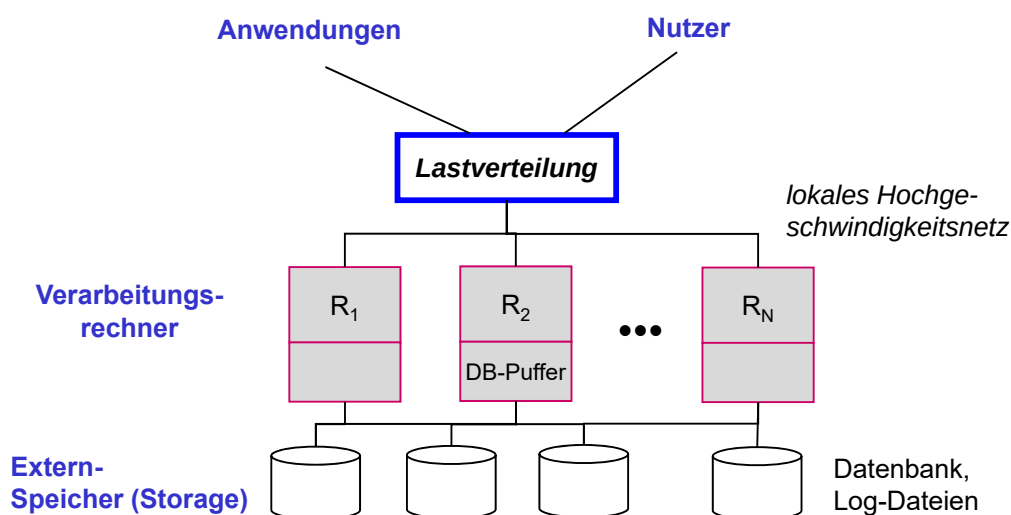


8. Shared-Storage DBS

- Einführung
- Synchronisation
 - zentrale Sperrverfahren
 - Schreib-, Leseauthorisierungen
 - verteilte Sperrverfahren
- Kohärenzkontrolle
 - Teilprobleme
 - On-Request-Invalidierung
- nahe Kopplung
- Beispiel-Implementierungen
 - IBM Parallel Sysplex
 - Azure SQL Hyperscale
 - Huawei GaussDB



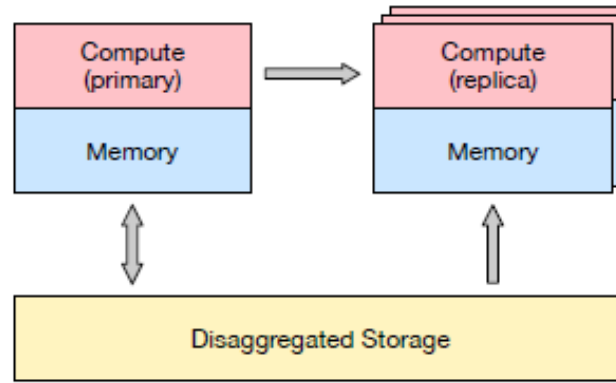
Grobaufbau Shared-Storage/Disk PDBS



- keine DB-Partitionierung unter Rechnern
 - Änderungen in der Rechneranzahl relativ einfach verkraftbar
 - hohes Potenzial zur Lastbalancierung und Intra-Query-Parallelisierung
- lokale Rechneranordnung erforderlich
- Externspeicherzugriff über Netzwerk (z.B. SAN)



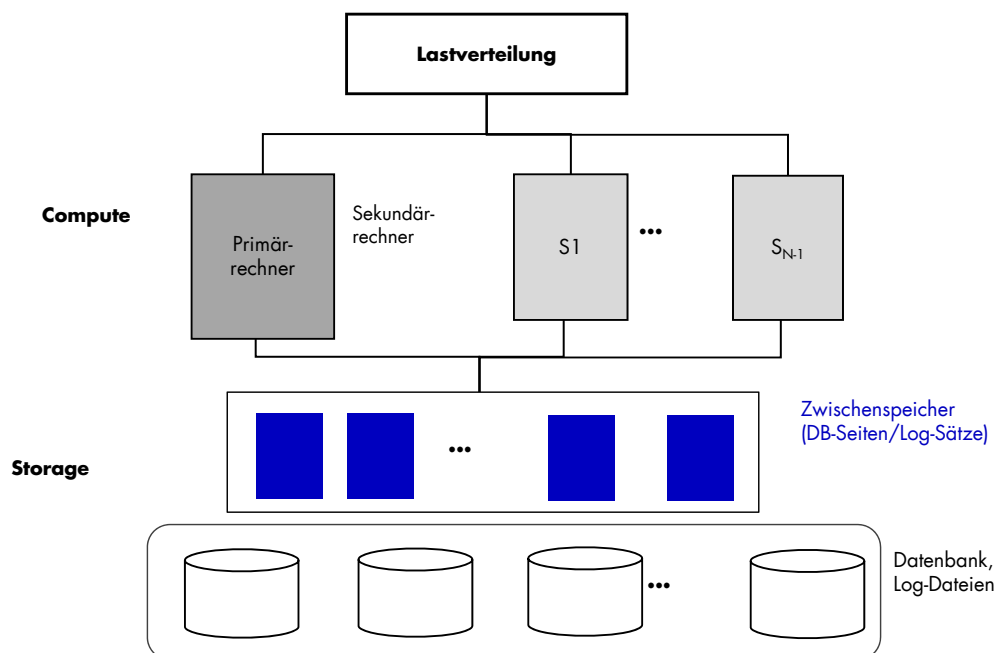
Shared Storage in Cloud-DBS



- Trennung Compute und Storage
 - bessere Skalierbarkeit/Kosten für viele Nutzer (tenants)
- initial Änderungen nur in 1 Knoten
 - Amazon Aurora, Azure SQL Hyperscale, Google Alloy
- neuere Systeme (Huawei GaussDB, Alibaba PolarDB) erlauben Änderungen in allen Knoten

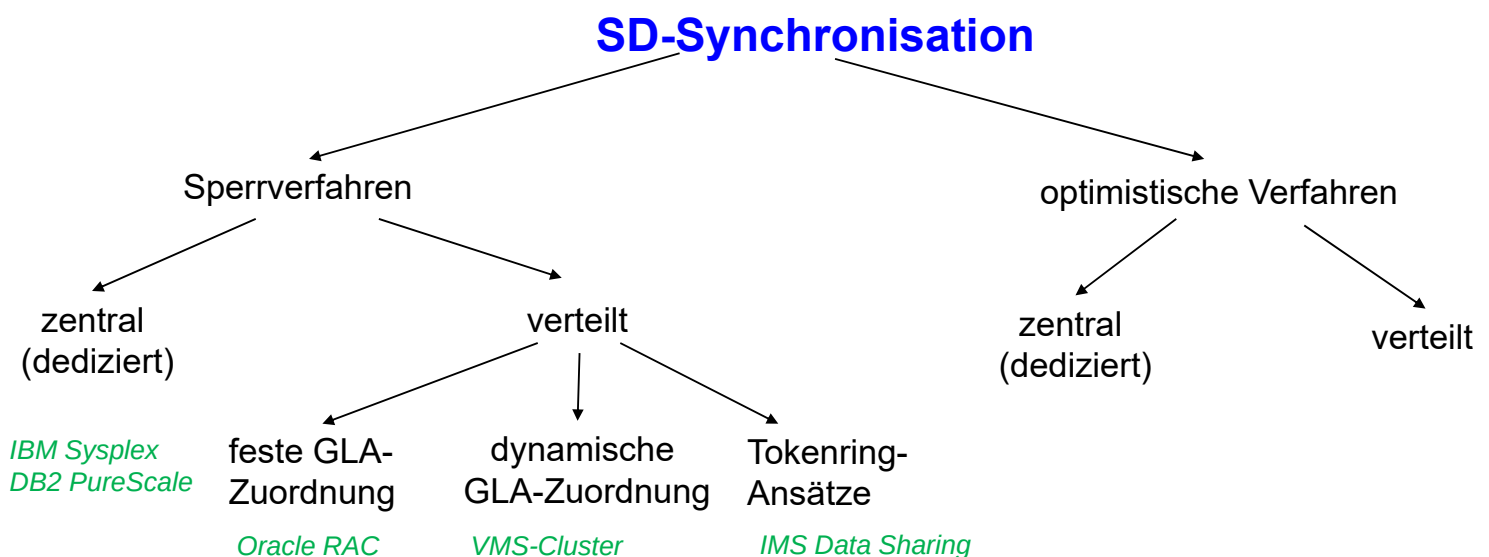


Cloud-natives Shared Storage DBS



Shared-Disk-DBS: Neue Probleme

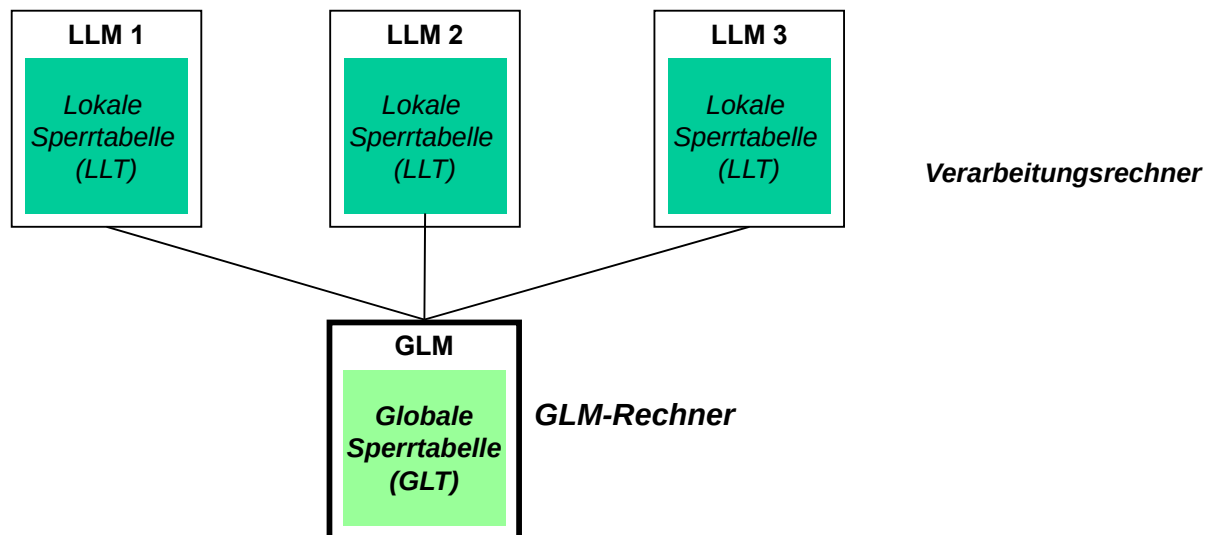
- **globale Synchronisation** (Concurrency Control)
 - Wahrung der globalen Serialisierbarkeit
 - möglichst wenige Synchronisationsnachrichten !
- **Kohärenzkontrolle**
 - dynamische Replikation von DB-Seiten im Hauptspeicher
 - Zugriff auf invalidierte Seiten zu vermeiden (mit geringem Aufwand)
- **Lastverteilung**
 - Lastbalancierung
 - Unterstützung rechnerpezifischer Lokalität
- **Logging und Recovery**
 - Crash-Recovery durch überlebende Rechner
 - Erstellung einer globalen Log-Datei durch Mischen lokaler Log-Daten
- **Nutzung einer nahen Kopplung**



- Sperrverfahren verwenden **globalen Lock-Manager (GLM)**
 - zu jedem Objekt hat genau einer der Rechner die globale Synchronisationsverantwortung (**GLA = Global Lock Authority**)
 - jeder Rechner hat **lokalen Lock-Manager (LLM)** für lokale Transaktionen



Zentrale Sperrverfahren



- zentrale Sperrverfahren: GLM für gesamte DB auf einem Rechner
 - Basisschema: 2 Nachrichten pro Sperranforderung
 - Nachrichtenbündelung erlaubt Reduzierung des Kommunikations-Overheads, verlängert jedoch die Antwortzeiten

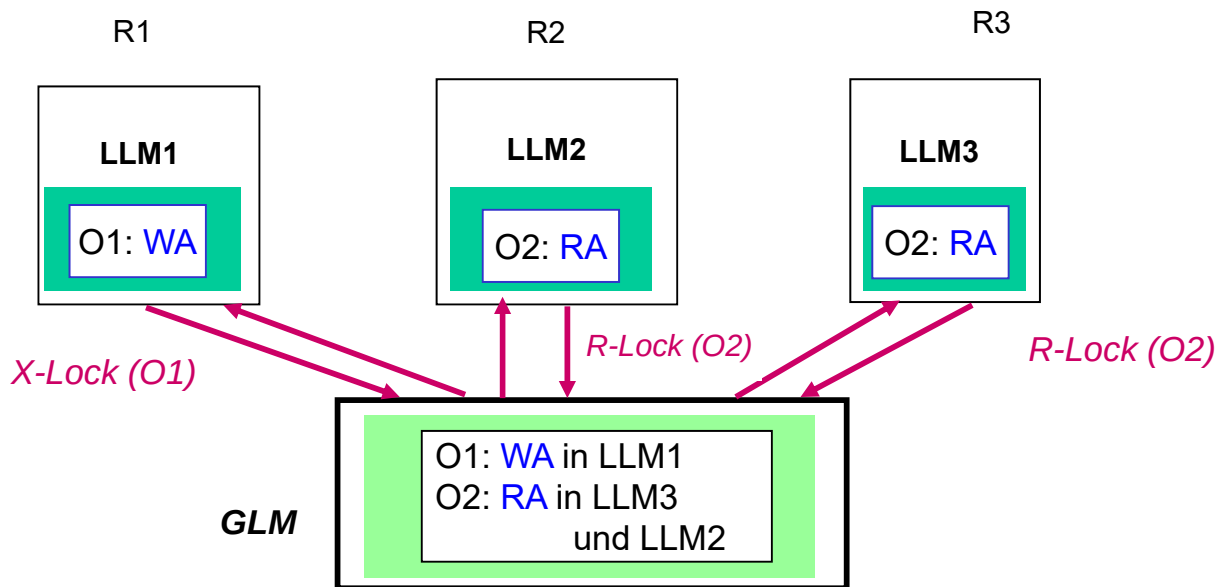


Schreib- und Leseautorisierungen

- Lokalität ermöglicht Einsparung globaler Sperranforderungen
- Vergabe von Schreib- und Leseautorisierungen an LLM
 - GLM erteilt bei Sperranforderung **Schreibautorisierung (write authorization WA)**, falls anfordernder Rechner allein "Interesse" an dem Objekt besitzt (*sole interest*)
 - bei Leseanforderung wird **Leseautorisierung (read authorization RA)** erteilt, sofern keine Schreibsanforderung vorliegt
- Schreib/Leseautorisierungen für lokale Gewährung / Freigabe von Sperren
=> **Einsparung von Nachrichten an GLM**
 - Schreibautorisierung für O: lokale Schreib- und Lesesperren für O durch LLM
 - Leseautorisierung unterstützt lokale Vergabe von Lesesperren
 - LLM behält Autorisierungen nach Transaktionsende
- Schreib/Leseautorisierungen verzögern Sperrzuteilungen an anderen Rechnern
 - vorheriger Entzug der Autorisierungen durch GLM
 - RA-Entzug nur für Änderungszugriffe



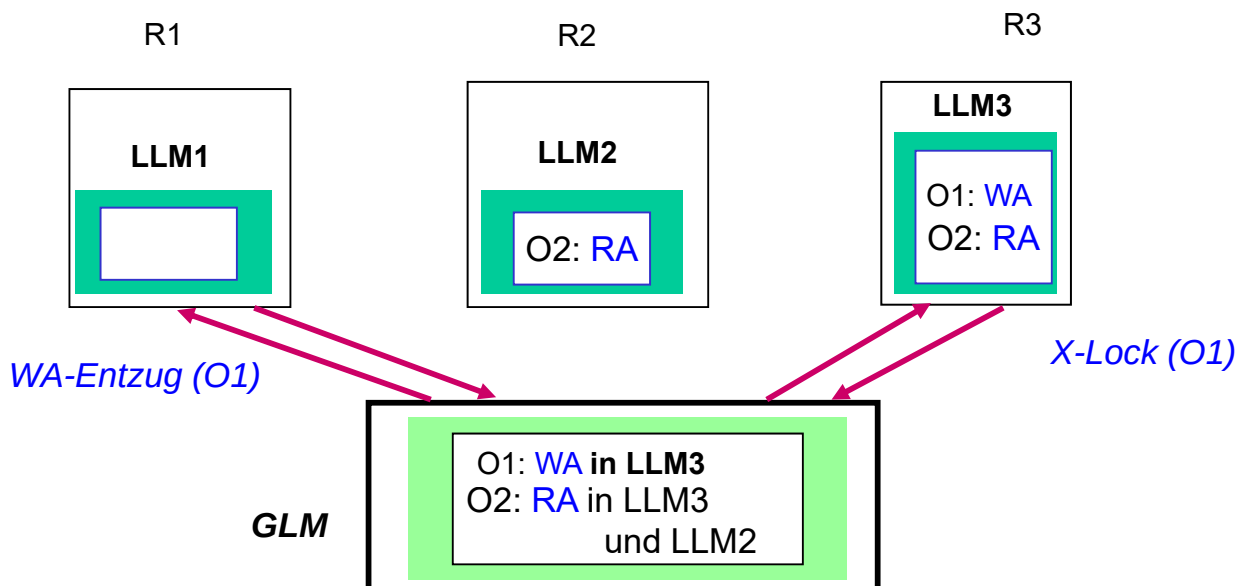
Einsatz von Autorisierungen



GLM = Global Lock Manager
 LLM = Local Lock Manager
 RA = Read Authorization
 WA = Write Authorization



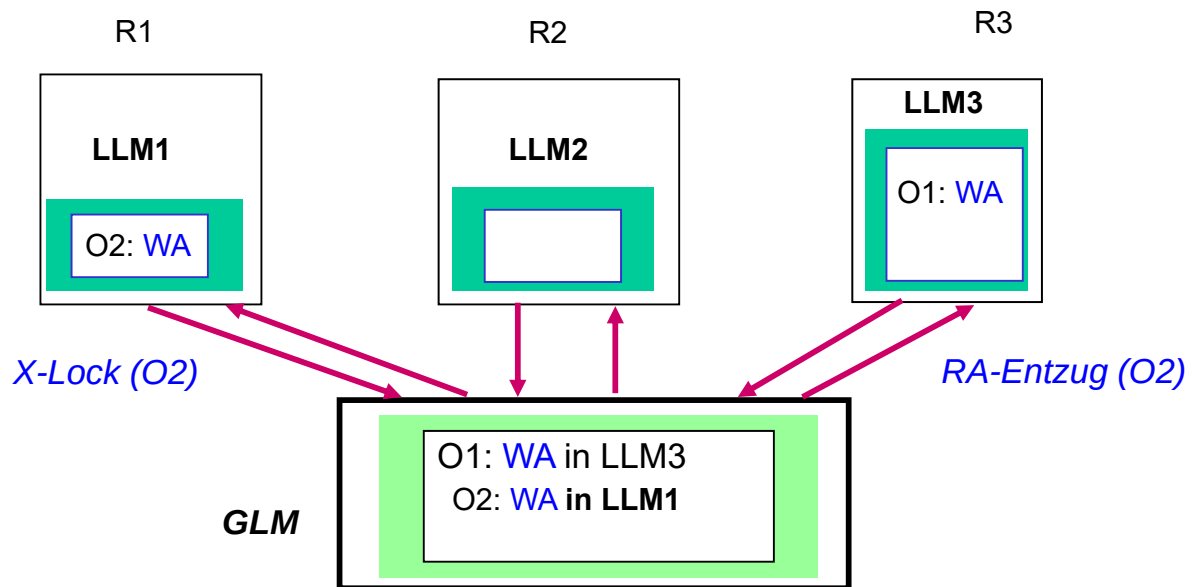
Einsatz von Autorisierungen: WA-Entzug



GLM = Global Lock Manager
 LLM = Local Lock Manager
 RA = Read Authorization
 WA = Write Authorization



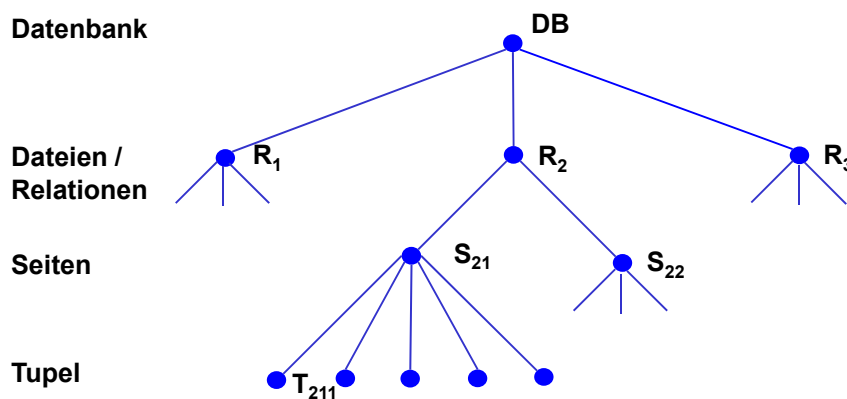
Einsatz von Autorisierungen: RA-Entzug



GLM = Global Lock Manager
 LLM = Local Lock Manager
 RA = Read Authorization
 WA = Write Authorization



Hierarchische Autorisierungen



- Autorisierungen können auf mehreren Ebenen der Objekthierarchie verwaltet werden
- Schreibautorisierung für Datei beinhaltet z. B. eine *implizite Schreibautorisierung* für alle zugehörigen Seiten und Sätze
 => vollkommen lokale Synchronisierung auf Datei
- Leseautorisierungen für Datei ermöglicht lokale Vergabe/Rückgabe von Lesesperren für alle Objekte der Datei



Verteilte Sperrverfahren

■ generelle Probleme zentraler Sperrverfahren:

- Engpass-Gefahr
- Verfügbarkeitsprobleme

=> Verteilung der Synchronisationsverantwortung auf mehrere Rechner

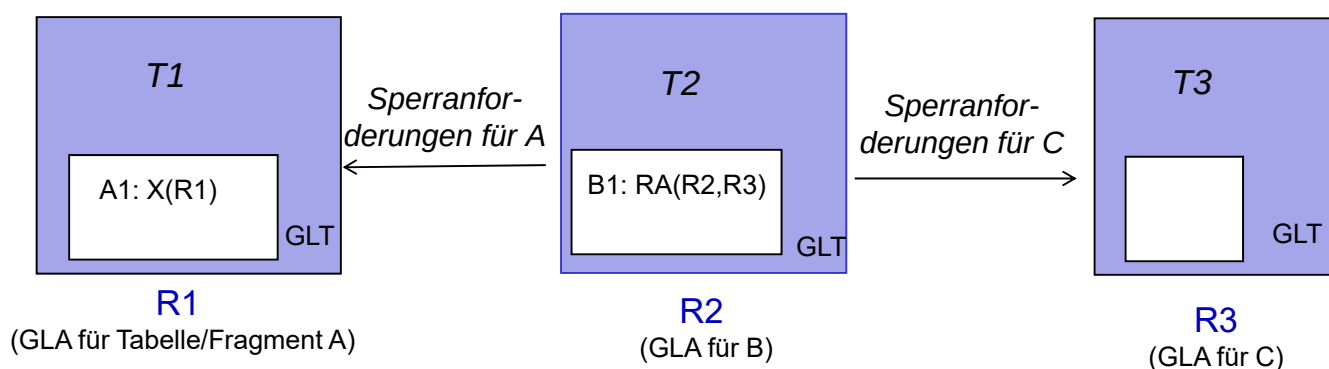
■ Unterfälle

- Synchronisation auf dedizierten Rechnern oder allgemeinen Verarbeitungsrechnern
- feste vs. dynamische Verteilung der GLA
- feste GLA-Partitionierung (Mapping Objekt-ID -> GLA-Rechner) analog zur Fragmentierung in Parallelen DBS, z.B.
 - Hash-Partitionierung oder
 - logische Fragmente (Tabellen, Tabellenbereiche)



Sperrverfahren mit logischer GLA-Verteilung

- Ziel: Koordinierung von GLA- und Lastverteilung zur Minimierung von Kommunikation
- lokale Sperrvergabe in Rechner R für Objekte, zu denen R GLA hat
 - stabiler als Schreibautorisationen
- zusätzliche Nutzung von Leseautorisationen sinnvoll



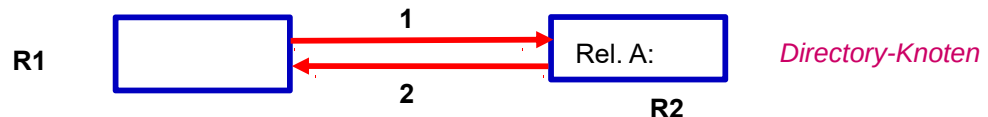
Dynamische GLA-Zuordnung

- Rechner, an dem ein Objekt zuerst referenziert wird, bekommt GLA für dieses Objekt
- Verwaltung der GLA-Zuordnung:
 - zentrale Tabelle
 - replizierte Tabelle
 - (hash-) partitionierte Tabelle
- zusätzliche Nachrichten zur Feststellung des GLM-Rechners
- GLA kann bei häufigen externen Sperranforderungen migrieren
- Beispielrealisierung: Distributed Lock Manager (DLM) für VMS-Cluster



Dynamische GLA-Zuordnung: Beispiel

1) Erste Sperranforderung für Relation A (in Rechner R1)



=> R1 erhält GLA für Relation A

alle weiteren Sperranforderungen in R1 bezüglich Relation A werden lokal synchronisiert

2) Sperranforderung auf Relation A in Rechner R3

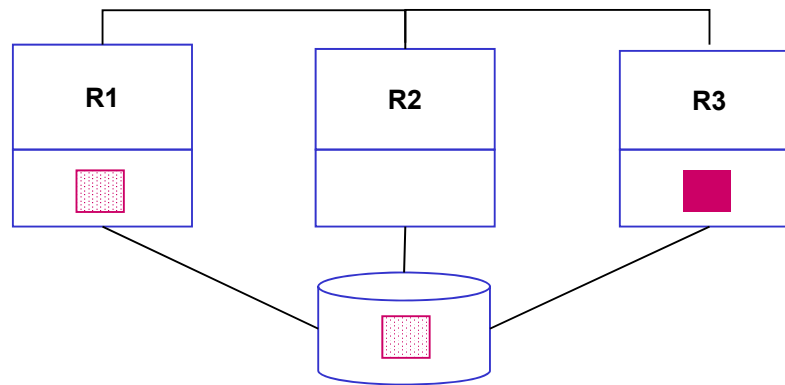


=> 4 Nachrichten für Sperre (R3 vermerkt sich, dass R1 GLM für A ist
(für Sperrkonversionen, Sperrfreigabe)

3) Freigabe der letzten Sperre auf A (optional)



Kohärenzkontrolle



■ Teilprobleme

1. Vermeidung oder Erkennung von Pufferinvalidierungen

- Broadcast-Invalidierung (nicht skalierbar)
- On-Request-Invalidierung
- Vermeidung ("Buffer Purge" bzw. Haltesperren)

2. Update-Propagierung: Bereitstellung der aktuellen Objektversionen

- Einfluss der Ausschreibstrategie (FORCE vs. NOFORCE)
- Seitentausch über Externspeicher vs. direkter Transfer von Seiten



Update-Propagierung bei NOFORCE

■ NOFORCE:

- geänderte Seiten werden bei Commit aus Performance-Gründen nicht auf Externspeicher durchgeschrieben
- Externspeicher-Version einer Seite potenziell veraltet

■ Einführung eines Page-Owners für geänderte Seiten

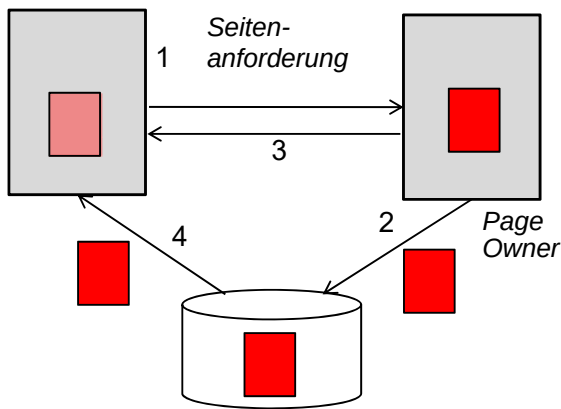
- stellt aktuelle Version einer Seite für anfordernde Rechner bereit
- i.a. Übertragung mittels Nachrichten
- schreibt geänderte Seite – asynchron – auf Externspeicher/Platte zurück: danach wird kein Page-Owner mehr benötigt

■ Realisierungs-Alternativen

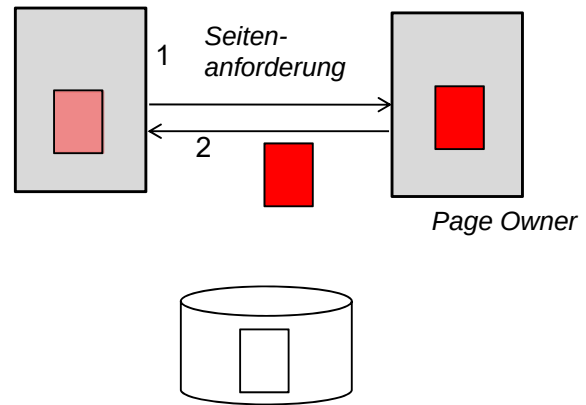
- *dynamische Page-Owner-Zuordnung*: Rechner, an dem die letzte Änderung erfolgt, wird zum Page-Owner
- *feste Page-Owner-Zuordnung*: Änderungen werden beim Unlock zum Page-Owner transferiert (falls die Änderung an einem anderem Rechner erfolgte)



Austausch von Seiten für SD



a) über Externspeicher (Disk oder SSD)



b) direkt

- Beschleunigung für a: nahe Kopplung
- Beschleunigung für b: *Remote Direct Memory Access* (RDMA)
 - Datentransferzeit von nur 1-3 μ s
- Cloud-native Systeme mit 1 Änderungsrechner
 - kein Austausch geänderter Seiten, sondern nur von Log-Daten mit denen alte DB-Seiten aktualisiert werden können



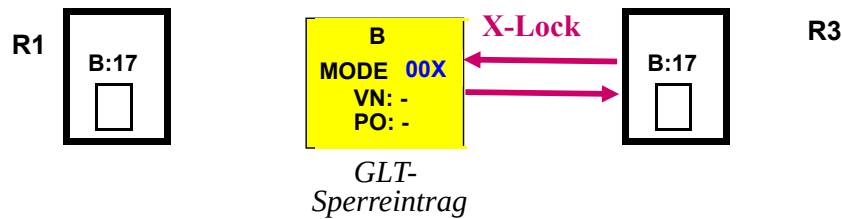
On-Request-Invalidierung

- Versionsnummer (Zähler) im Seitenkopf u. globaler Sperrtabelle
 - bei jeder Änderung: Erhöhung der Nummer in Seite und Ablage des aktuellen Wertes in der globalen Sperrtabelle
 - vor globaler Sperranforderung wird überprüft, ob Seite im lokalen Puffer vorliegt und mit welcher Versionsnummer
 - Invalidierung wird bei Sperrbearbeitung über Versionsnummern festgestellt
- globale Sperrtabelle führt auch Page-Owner (bei dynamischer Page-Owner-Zuordnung)
- weitgehende Einsparung von Nachrichten zur Kohärenzkontrolle
 - Anpassung der erweiterten Sperrinfo bei Freigabe von Schreibsperrern (Unlock)
 - Erkennung von Pufferinvalidierungen auf Seitenebene bei Lock-Request („on request“)
 - Bestimmung des Page-Owners bei Lock-Request
 - Nachrichten nur für Page-Requests, jedoch i.a. schneller als disk I/O

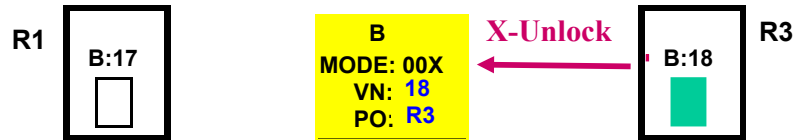


On-Request-Invalidierung: Beispiel

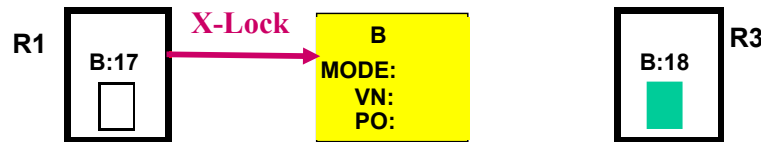
1.) erste Änderung von Seite B in R3



2.) Situation nach Freigabe der Schreibsperre durch R3



3.) Bei Sperranforderung durch R1 wird Pufferinvalidierung erkannt

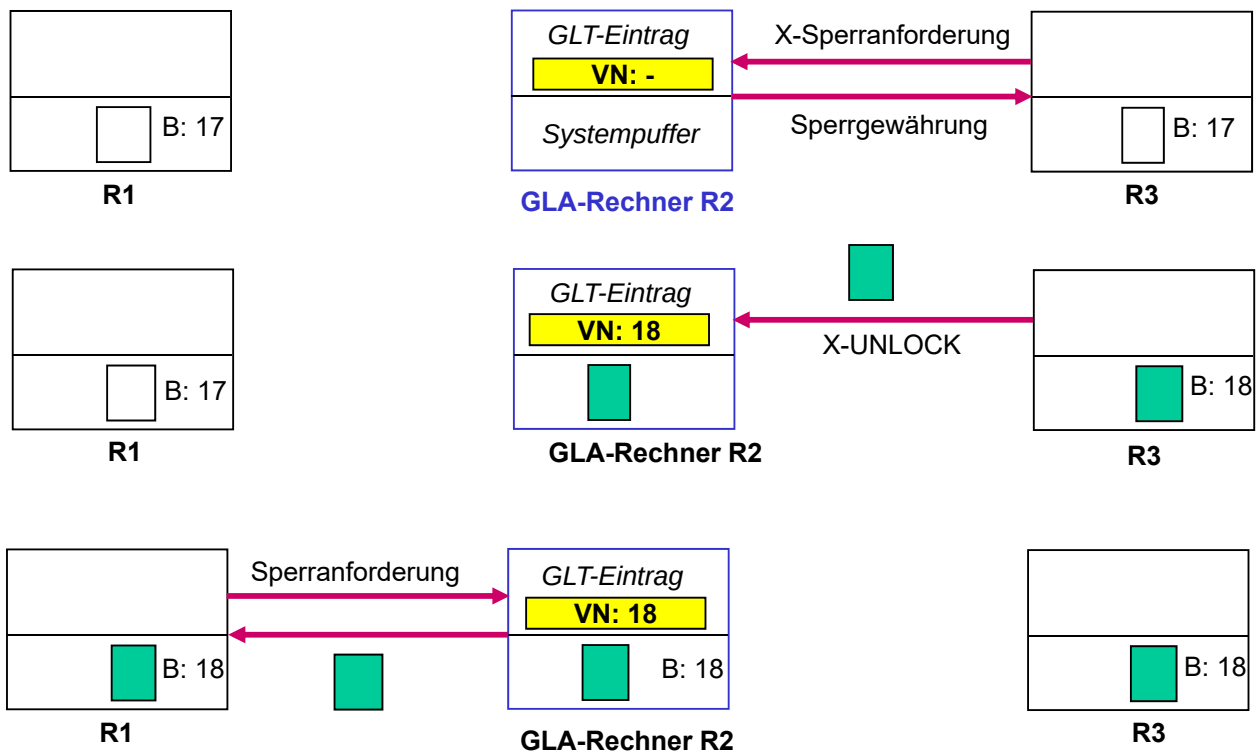


Feste Page-Owner-Zuordnung

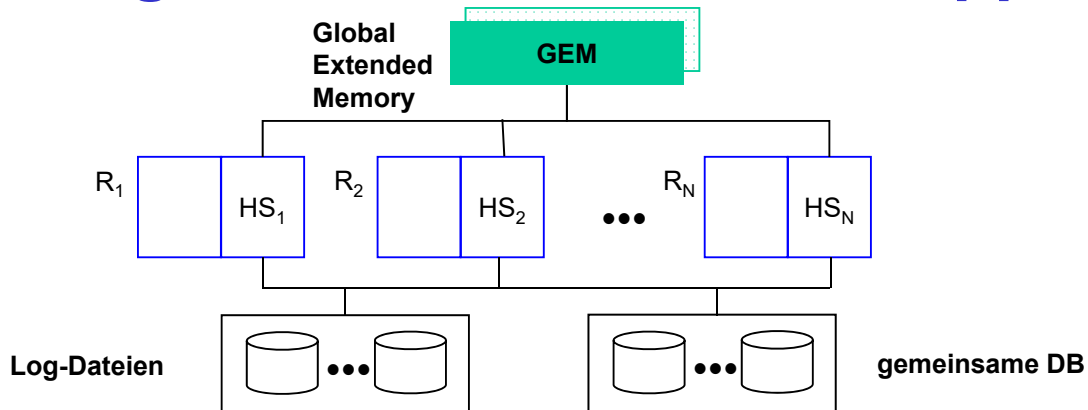
- zu jeder Seite (Partition) existiert ein Rechner, dem vorab die „Page Ownership (PO)“ zugeordnet wurde
 - alle Änderungen sind zum Page-Owner zu transferieren
- starke Kommunikationseinsparungen durch gemeinsame Zuordnung von GLA und Page-Ownerships
 - GLA und PO zu einem Objekt am gleichen Rechner
 - Transfer geänderter Seiten bei EOT zusammen mit X-Sperrfreigabe
 - Kombination der *Page Requests* mit *Lock Requests*
 - oft Seitenbereitstellung mit Sperrgewährung
 - keine zusätzliche Nachrichten zur Kohärenzkontrolle (Erkennung von Invalidierungen und Update-Propagierung)
- sinnvoll v.a. bei logischer GLA-Zuordnung
 - globale Synchronisation u. Seitenübertragungen auf alle Rechner verteilt
 - lokale GLA erspart Page-Transfers zum/vom Owner
 - Koordinierung von GLA- und Lastzuordnung reduziert #Seitenübertragungen



Kombinierte GLA- und Page-Owner-Zuordnung („Primary-Copy-Ansatz“)



Nutzung einer nahen Rechnerkopplung



■ gemeinsame Halbleiterspeicherbereiche:

- beschleunigte Kommunikation (Austausch geänderter Seiten)
- globale Datenstrukturen zur Synchronisation (Sperrtabelle) und Lastverteilung
- globaler Systempuffer und Allokation von DB-Dateien
- Logging (lokale und globale Log-Dateien)

■ Speichereigenschaften:

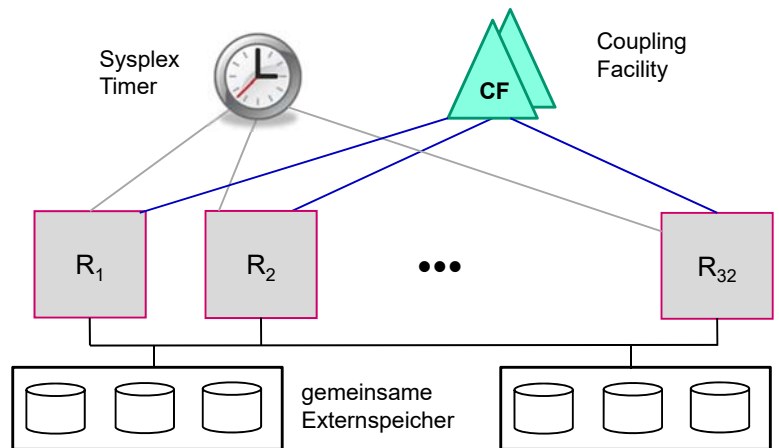
- schneller, synchroner Zugriff (< 10 µs)

■ zunehmende Nutzung in cloud-nativen Systemen mit disaggregated Memory



IBM Parallel Sysplex

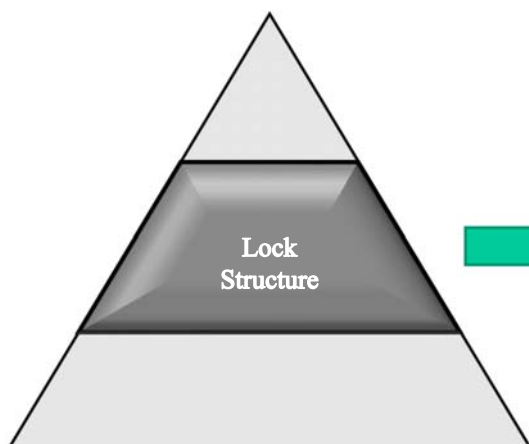
- Shared-Disk-Architektur für zOS-Umgebungen (DB2, IMS, Adabas ...)
 - Unix: DB2 PureScale mit RDMA-Kommunikation
- nahe Kopplung von IBM-Großrechnern über **Coupling Facility (CF)** und **Sysplex Timer**
- CF: Spezialprozessor für globale Kontrollaufgaben:
 - Sperrbereich (lock structure)
 - Pufferbereich (cache structure)
 - Listenbereich (list structure)
- dynamische Lastbalancierung („Global Workload Manager“)



Quelle: IBM

Sysplex Global Locking

- globales Sperrinteresse auf Rechnerebene wird in CF verwaltet
 - reduzierte Sperrinfo auf Ebene von Hash-Klassen
 - entspricht grob Schreib- bzw. Leseautorisationen
 - Transaktionssperren / Wartebeziehungen bei LLMs (IRLM = Internal Resource Lock Manager) der Knoten
- Kommunikation zwischen Rechnern nur für (seltene) Konfliktfälle



Lock Table (Hash Table)

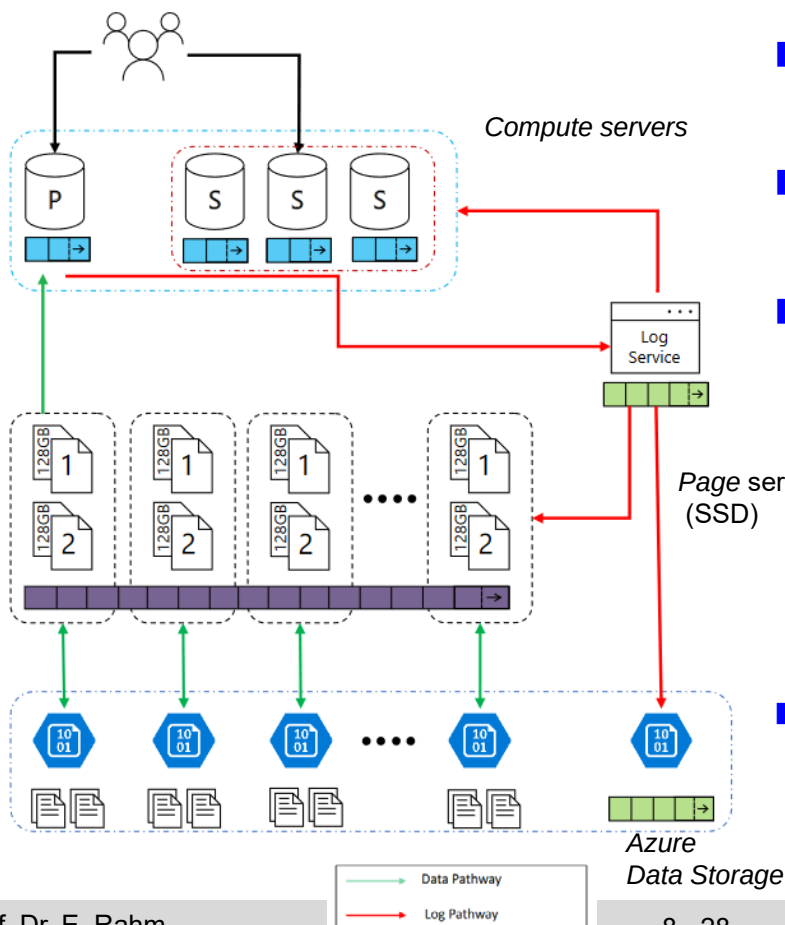
Exclusive	Shared Lock Status							
00	1	1	0	0	0	0	0	0
02	1	0	0	0	0	0	0	0
...								
...								
01	0	0	0	0	0	0	0	0

Sysplex Kohärenzkontrolle

- FORCE in globalen Puffer der CF (ca. 20 μ s pro Seite)
- Multicast-Invalidierung
 - CF kennt Pufferstatus jedes Knotens
 - für in CF geschriebene geänderte Seite werden hardware-gestützt Rechner mit betreffender Seite über Invalidierung informiert
 - Bitvektor pro Rechner zeigt pro Pufferrahmen an, ob Invalidierung vorliegt
- Lesen der aktuellen Seiten bei Invalidierung von CF-Puffer bzw. Externspeicher
- asynchrones „Cast-Out“ geänderter Seiten vom CF-Puffer in Externspeicher



Azure SQL Hyperscale

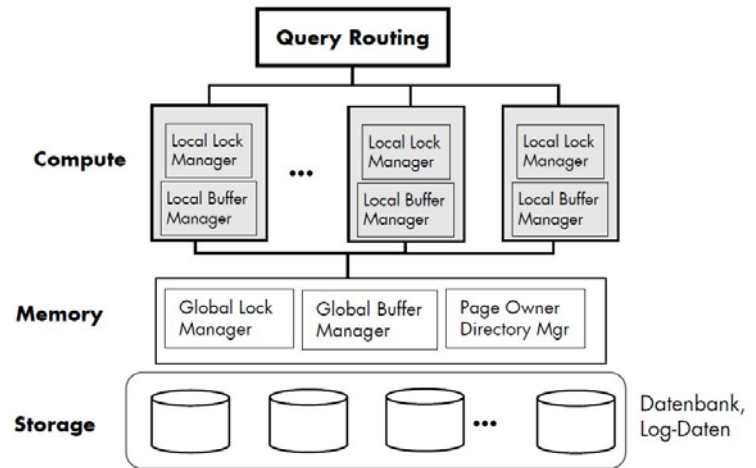


- 1 Primary und n (read-only) Secondaries
- Unterstützung für MySQL und PostgreSQL
- nur Log-Daten werden ausgeschrieben
 - stark reduzierter I/O_Aufwand
 - „Log is the database“
 - Rekonstruktion geänderter Seiten in Page Servern
 - Secondaries nutzen Log-Daten zur Aktualisierung gepufferter Seiten
- fehlertoleranter Log-Service



Huawei GaussDB*

- Änderungen in jedem Knoten
- nahe Kopplung mit RDMA-Zugriff auf gemeinsamen Hauptspeicher
 - globale Daten über Hashing auf mehrere Memory-Knoten verteilt
- Autorisierungen und Page Owner-Aufgaben in Compute-Knoten
 - lokale PO erlaubt lokales Sperren
- affinitätsbasiertes Query-Routing
 - Queries sollen an Knoten mit PO für Mehrzahl benötigter Objekte gehen
 - Nutzung eines Referenzgraphen (welche Objekte werden in Queries zusammen verwendet) sowie von Machine Learning zur Bestimmung von für Anfrage wahrscheinlich benötigter Daten



* G. Li et al.: GaussDB: A Cloud-Native Multi-Primary Database with Compute-Memory-Storage Disaggregation. PVLDB 2024



Zusammenfassung

- Optimierungsziele
 - minimale Zahl von Nachrichten für Synchronisation u. Kohärenzkontrolle
 - koordinierte Lösung von Synchronisation und Kohärenzkontrolle
- Nutzung von Lokalität zur Einsparung externer Sperranforderungen
 - Nutzung einer lokalen GLA
 - Schreibautorisationen und Leseautorisationen
 - hierarchischer Sperransatz
- Kohärenzkontrolle
 - kombinierbar mit Synchronisation (z.B. On-Request-Invalidierung)
 - Update-Propagierung: NOFORCE + direkter Seitenaustausch zwischen Rechnern
 - affinitätsbasierte Lastverteilung (Lokalität) begrenzt Pufferinvalidierungen und Seitentransfers
- nahe Kopplung kann effektiv genutzt werden
 - Variante: RDMA/CLX für schnellen Zugriff auf entfernte Hauptspeicher
 - reduzierte Notwendigkeit einer hohen Lokalität
- Cloud DBS: Änderungen nur in 1 Rechner oder allgemeine Lösung mit naher Kopplung

