UNIVERSITÄT LEIPZIG

Mainframe Internet Integration

Prof. Dr. Martin Bogdan
Prof. Dr.-Ing. Wilhelm G. Spruth

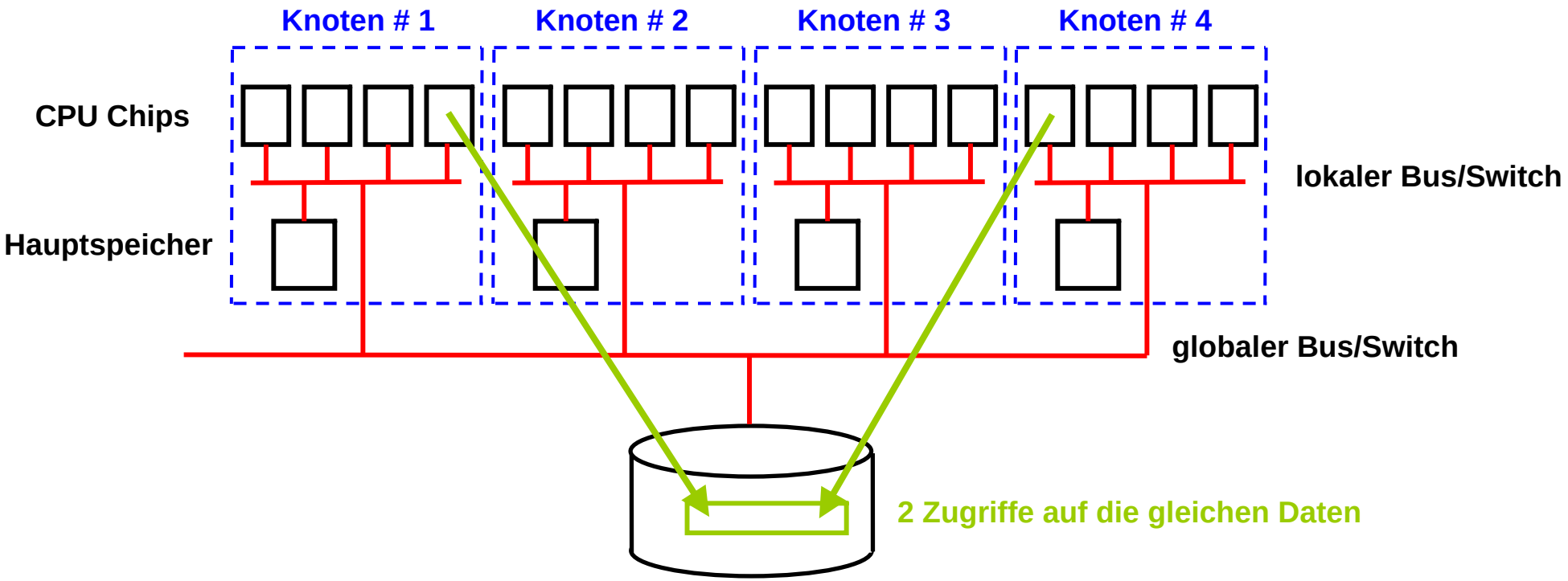
SS2013

Sysplex Teil 2

Coupling Facility

zSeries Coupling Facility

Großrechner bearbeiten mehrere 1000 Transaktionen / Sekunde. Kritisch ist die Einhaltung der ACID Bedingungen.



Cluster haben den Vorteil, dass mehrfache Instanzen (Kopien) des Betriebssystems vorhanden sind. Deshalb bestehen alle modernen Großrechner aus einem Cluster von SMPs.

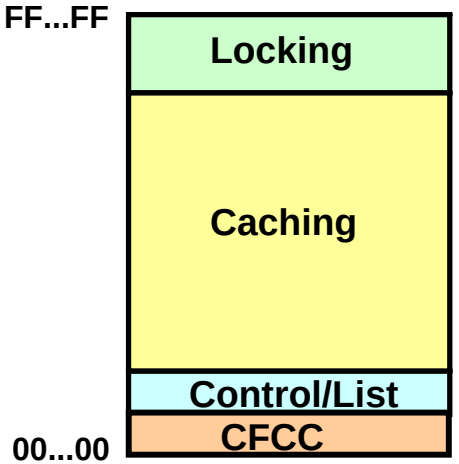
Allerdings besteht hierbei das Problem, den gleichzeitigen Zugriff mehrerer Knoten auf den gleichen Datenbestand zu lösen. Dies geschieht mit Hilfe von Locks (Sperren) auf Teile der gemeinsam genutzten Daten. Die Effektivität des Lock Managements (Sperrverwaltung) bestimmt die Skalierungseigenschaften eines Clusters.

Coupling Facility

Die Coupling (CF) Facility ist in Wirklichkeit ein weiterer System z Rechner mit spezieller Software. Die Aufgaben der CF sind:

- **Locking**
- **Caching**
- **Control/List Structure Management**

Der Hauptspeicher der Coupling Facility enthält einen als „Coupling Facility Control Code“ (CFCC) bezeichnetes Betriebssystem, sowie Speicherbereiche für Locking, Caching und Control/List Strukturen.



Die Coupling Facility ist über Glasfaser Verbindungen mit einem optimierten Protokoll (Infiniband) und spezieller Hardware Unterstützung mit den Knoten (Systemen) des Sysplex verbunden.

Coupling Facility (CF)

Die wichtigste Aufgabe der Coupling Facility ist ein zentrales Lock Management für die angeschlossenen Knoten. Der zentrale Lock Manager des SAP System R/3 hat in Ansätzen eine ähnliche Funktionalität, jedoch ohne die CF spezifischen Eigenschaften.

Der größte Teil des Hauptspeichers der Coupling Facility wird als Plattenspeicher Cache genutzt. Der CF Cache dupliziert den Plattenspeicher Cache (Buffer Pool) in den einzelnen Systemen. Ein Cast out der CF Cache auf einen Plattenspeicher erfolgt über ein System (Knoten).

CF Cache Cross-Invalidate (ungültig machen) Nachrichten gehen nur an die betroffenen Systeme.

Control und List Strukturen dienen der Sysplex Cluster weiten Verwaltung. Ein Beispiel ist ein den ganzen Cluster umfassendes RACF Sicherheits Subsystem. Ein weiteres Beispiel ist der bereits erwähnte WebSphere MQ Cluster (siehe Wintersemester MQSeries, Teil 4).

Eine interessante Überlegung: Das SAP System R/3 würde erheblich vom Vorhandensein einer Coupling Facility profitieren. Das Problem ist: SAP System R/3 ist als Multi-Plattform Software ausgelegt, lauffähig nicht nur unter z/OS, sondern auch unter den verschiedenen Unix Dialekten, Linux und Windows. Um eine CF nutzen zu können, wären identische Hardware Einrichtungen auf Sparc, Itanium, PowerPC und x86 Rechnern erforderlich.

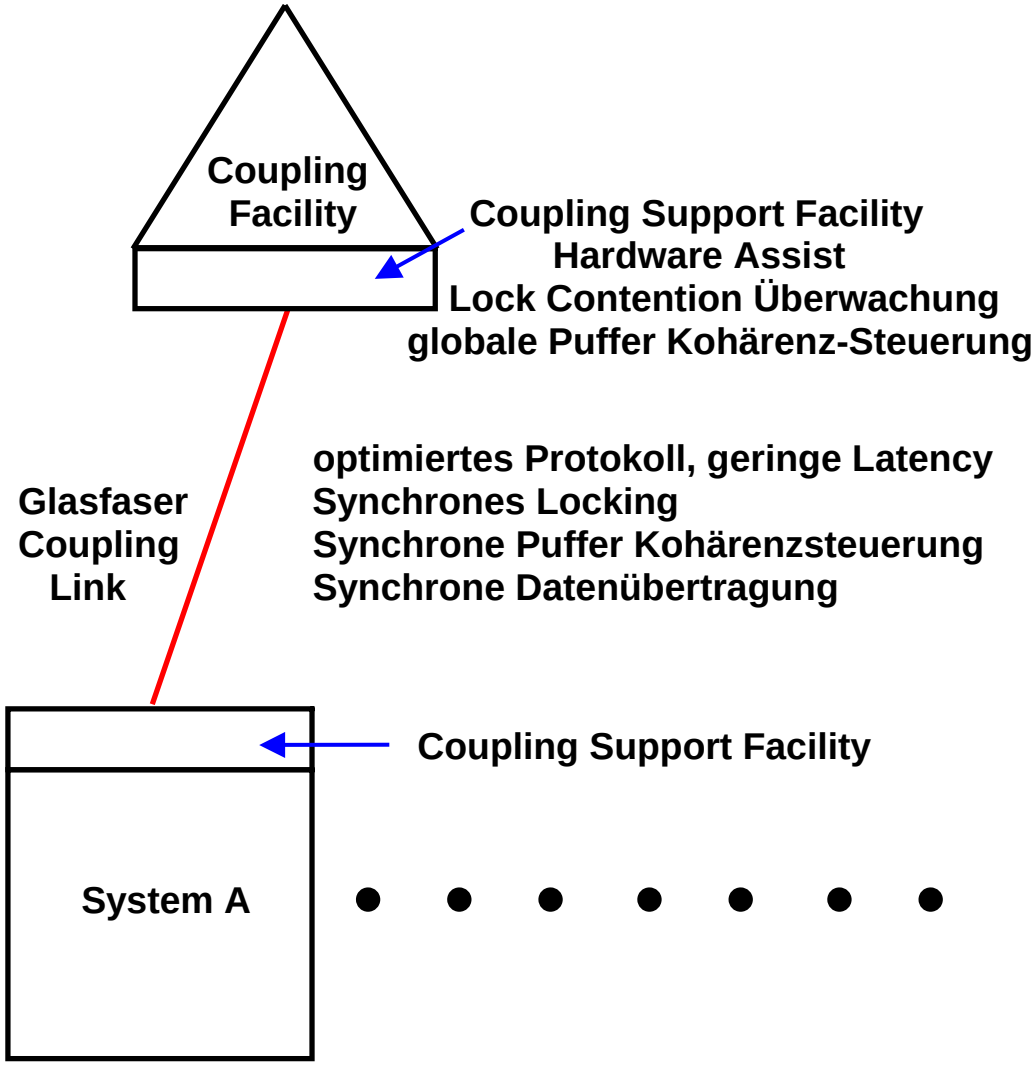
Die Lauffähigkeit auf unterschiedlichen Platformen hat ihren Preis.

Coupling Support Facility

Jeder Knoten (System) eines Sysplex sowie die CF selbst enthalten eine spezielle Hardware Einrichtung, die „Coupling Support Facility“. Diese bewerkstelligt die Kommunikation und den Datenaustausch zwischen Knoten und CF.

Die Coupling Support Facility besteht aus je einer Glasfaserverbindung (800 MByte/s theoretisches Maximum), einem dedizierten Link-Prozessor für die Verarbeitung des (Infiniband) Übertragungsprotokolls und einer Erweiterung des System z Maschinenbefehlssatzes. Die zusätzlichen Maschinenbefehle realisieren die nahe Kopplung zwischen Knoten und CF und ermöglichen allen Knoten den Zugriff auf die Inhalte in den CF-Datenstrukturen.

Das Kommunikationsprotokoll wurde für eine besonders geringe Latenzzeit optimiert, welches *synchrone Zugriffe* eines jeden Knotens auf die CF ermöglicht. Synchron bedeutet, dass ein auf einem Knoten laufender Prozess während eines Zugriffs auf die Coupling Facility blockiert: Er bleibt im Zustand „running“, und es findet kein Prozesswechsel und kein Übergang "User-Status - Kernel-Status" statt. Die Zugriffsgeschwindigkeit liegt im Mikrosekundenbereich und ist somit weitaus effizienter als die Kommunikation über allgemeine Protokolle bei loser Rechnerkopplung.



Die Coupling Facility ist durch eine Punkt-zu-Punkt Glasfaserleitung mit jedem System (Knoten) des Sysplex verbunfen (Coupling Facility Link).

Es wird ein spezielles Verbindung-Protokoll (Infiniband) mit besonders geringer Latency eingesesetzt .

Die CF Glasfaser Verbindung wird durch spezielle Hardware Einrichtungen und durch zusätzliche Maschinenbefehle in jedem angeschlossenen System unterstützt (Coupling Support Facility).

Die Coupling Facility kann in einem angeschlossenen Rechner ohne Unterbrechung des laufenden Prozesses Daten in spezielle Speicherbereichen (Bit Vektoren) abändern.

Anbindung eines Systems an die Coupling Facility

Komponenten einer Coupling Facility

Die nahe Kopplung über die CF wird für die leistungskritischen Kontrollaufgaben in Shared-Disk-Clustern genutzt, um durch die effiziente Realisierung eine hohe Skalierbarkeit zu erreichen. Dies betrifft die globale Synchronisation über Sperrverfahren (Locking), die Kohärenzkontrolle der Pufferinhalte mit schnellem Austausch geänderter Daten zwischen den Knoten sowie die flexible Lastverteilung. Entsprechend werden spezifische Funktionen (Maschinenbefehle) für Locking, Caching und Queuing sowie zugeschnittene Datenstrukturen im Hauptspeicher der CF bereitgestellt. Die Hauptspeicher-Ressourcen der CF-können dynamisch partitioniert und einer der CF-Strukturen zugewiesen werden. Innerhalb derselben CF sind mehrere CF-Strukturen desselben oder unterschiedlichen Typs möglich. Die auf diesen Strukturen bereitgestellten Funktionen bzw. Cluster-Protokolle repräsentieren drei Verhaltensmodelle:

Host hardware: Wenn eine CPU eine Anforderung an die Coupling Facility startet, wird diese zunächst mittels Firmware (Microcode, erläutert später im Modul Virtualisierung Teil 3) und spezieller Hardware verarbeitet, ehe sie über das Coupling Link weitergereicht wird.

Coupling Link: Die Anforderung wird dann über das Coupling Link an die Coupling Facility übertragen. Die Übertragungszeit wird durch die Link Geschwindigkeit, die Datenmenge, die Distanz zur Coupling Facility (Lichtgeschwindigkeit) bestimmt. Die Übertragung erfolgt nach Möglichkeit synchron, d.h. der laufende Prozess wird nicht unterbrochen. Deshalb ist die Latenz (die Zeit, die für die Übertragung eines einzelnen Bytes benötigt wird) besonders kritisch. Um die Latenzy zu minimisieren, verwenden CF Links deshalb entweder ein spezielles CF Link Protokoll, oder das Infiniband Protokoll. Beide zeichnen sich durch eine besonders geringe Latency aus.

Coupling facility hardware: Die eintreffende Anforderung wird durch die Hardware und den Code in der Coupling Facility (Coupling Facility Control Code, CFCC) bearbeitet. Je nach Bedarf sendet die Coupling Facility Nachrichten an die Knoten Rechner des Sysplex zurück.

Coupling Facility Control Code

Wir hatten im Zusammenhang mit der Hardware System Area (HSA), den System Assist Prozessoren (SAP), dem Channel Subsystem und dem PR/SM Hypervisor den Begriff „Firmware“ erläutert.

Eine Coupling Facility ist ein normaler System z Rechner, der Coupling Facility Control Code (CFCC) ausführt. Es handelt sich hierbei um ein normales hochspezialisiertes Betriebssystem mit einigen Besonderheiten. CFCC ist jedoch Firmware – erläutert im Modul Virtualisierung Teil 3. CFCC Code ist aber gleichzeitig normaler System z Maschinencode.

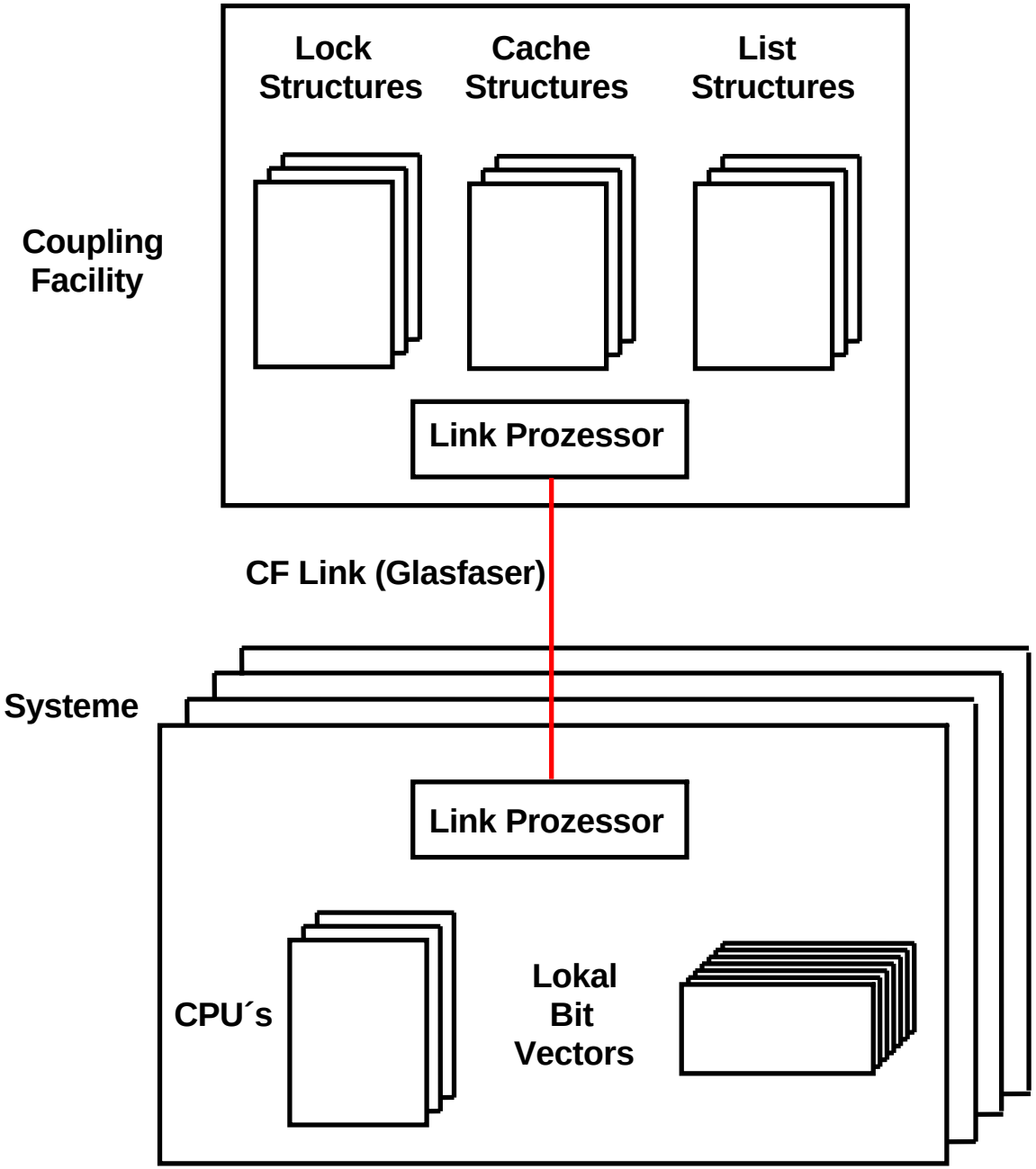
Dies bedeutet, eine Coupling Facility verhält sich wie eine Black Box. z/OS kann der CF eine Nachricht schicken, und die CF reagiert irgendwie darauf. Weder der z/OS Systemprogrammierer, und erst Recht nicht der Anwendungs-programmierer sind in der Lage, die CF zu programmieren.

Es ist deshalb möglich, für Experimentierzwecke CFCC Code in einer virtuellen Maschine unter z/VM laufen zu lassen. Dies haben wir auf dem Mainframe Rechner unseres Lehrstuhls installiert. Es wurde ein vollständiger virtueller Sysplex eingerichtet, der für Ausbildungs- und Übungszwecke genutzt werden kann.

Coupling Facility Strukturen

Die nahe Kopplung über die CF wird für die leistungskritischen Steuerungslaufgaben genutzt, um durch die effiziente Realisierung eine hohe Skalierbarkeit zu erreichen. Dies betrifft die globale Synchronisation über Sperrverfahren (Locking), die Kohärenzkontrolle der Pufferinhalte mit schnellem Austausch geänderter Daten zwischen den Knoten sowie die flexible Lastverteilung. Entsprechend werden spezifische Funktionen (Maschinenbefehle) für Locking, Caching und Queuing sowie zugeschnittene Datenstrukturen im Hauptspeicher der CF bereitgestellt. Die CF Hauptspeicher-Ressourcen können dynamisch partitioniert und einer der CF-Strukturen zugewiesen werden. Innerhalb derselben CF sind mehrere CF-Strukturen desselben oder unterschiedlichen Typs möglich. Die auf diesen Strukturen bereitgestellten Funktionen bzw. Cluster-Protokolle repräsentieren drei Verhaltensmodelle:

- **Lock-Modell: Es unterstützt feingranulares, globales Locking für hohe Transaktionsverarbeitungs-Performance und eine Signalisierung von Zugriffskonflikten.**
- **Cache-Modell: Es liefert eine globale Kohärenz-Steuerung für die verteilten Pufferinhalte der einzelnen Knoten sowie einen globalen Puffer in der CF (Shared Data Cache).**
- **List Modell (Queue-Modell): Es implementiert einen umfangreichen Satz an Listen und Queuing-Konstrukten für die Verteilung von Arbeitslasten, zur Realisierung einer schnellen Nachrichtenübertragung (Message Passing) sowie für die Verwaltung globaler Status- Informationen. Ein Beispiel für letzteres ist die Verwaltung globaler Sicherheitsrechte.**



Die Coupling Facility unterhält in ihrem Hauptspeicher getrennte Strukturen für die Verwaltung von

- **Locks (Sperren)**
- **Sysplex weiter Daten Cache**
- **Listen für Sysplex-weite Aufgaben.**

Die einzelnen Systeme des Sysplex sind mit jeder dieser Strukturen über das CF Link und die Link Prozessoren (Teil der Coupling Support Facility) direkt verbunden. Die Kommunikation CPU – CF wird durch spezifische Maschinen-befehle unterstützt.

Jedes System unterhält lokale, als „Bit Vectoren“ bezeichnete Speicher-bereiche für eine logische Verbindung zu jeder Struktur in der CF. Es existieren getrennte Bit Vektoren für die Lock Strukturen, Cache Strukturen und List (Queue) Strukturen.

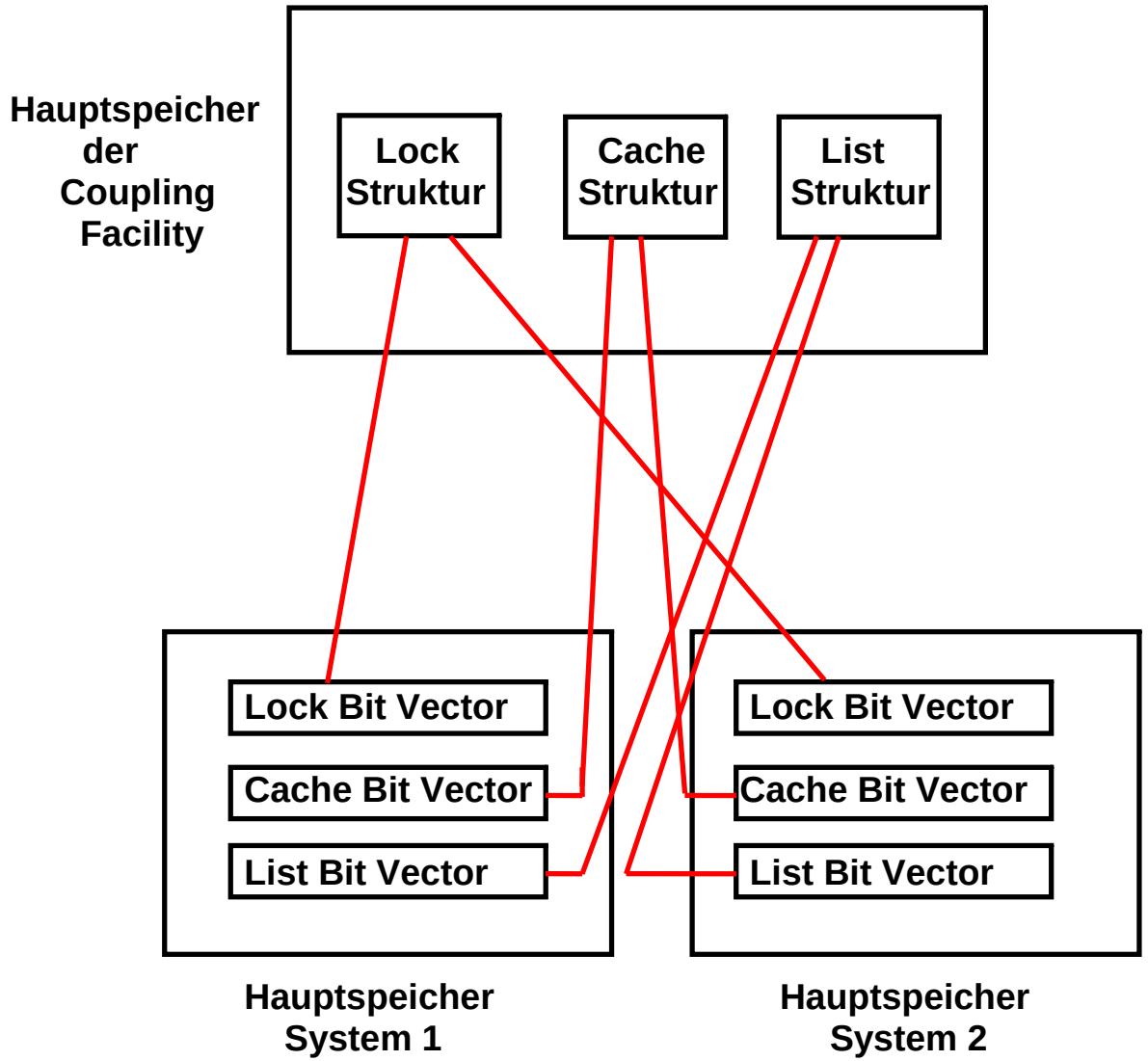
Zuordnung von Bit Vektoren zu CF Strukturen

Die folgende Abbildung zeigt die Anbindung der Knoten (Systeme) an die CF und die dort verwalteten globalen Datenstrukturen.

Jeder angeschlossene Rechner dupliziert in seinem Hauptspeicher eine Untermenge der in der Coupling Facility gespeicherten Daten. Diese Untermengen werden als Bit Vektoren bezeichnet.

Der Zugriff erfolgt dabei nicht nur von den Systemen (Knoten) auf die CF, sondern die CF kann auch direkt (ohne Involvierung des Betriebssystems) auf bestimmte Hauptspeicherinhalte der Systeme, spezifisch die Bit-Vektoren, zugreifen und ändern. Solche Bit-Vektoren existieren für jede logische Verbindung zu einer CF-Datenstruktur und erlauben z.B. die schnelle Signalisierung von Zugriffskonflikten (s.u.).

In den nächsten Abschnitt betrachten wir die Nutzung der CF-Lock- und Cache-Strukturen zur Realisierung der clusterweiten Synchronisation (Locking) und Kohärenzkontrolle.



Zuordnung von Bit Vektoren zu CF Strukturen

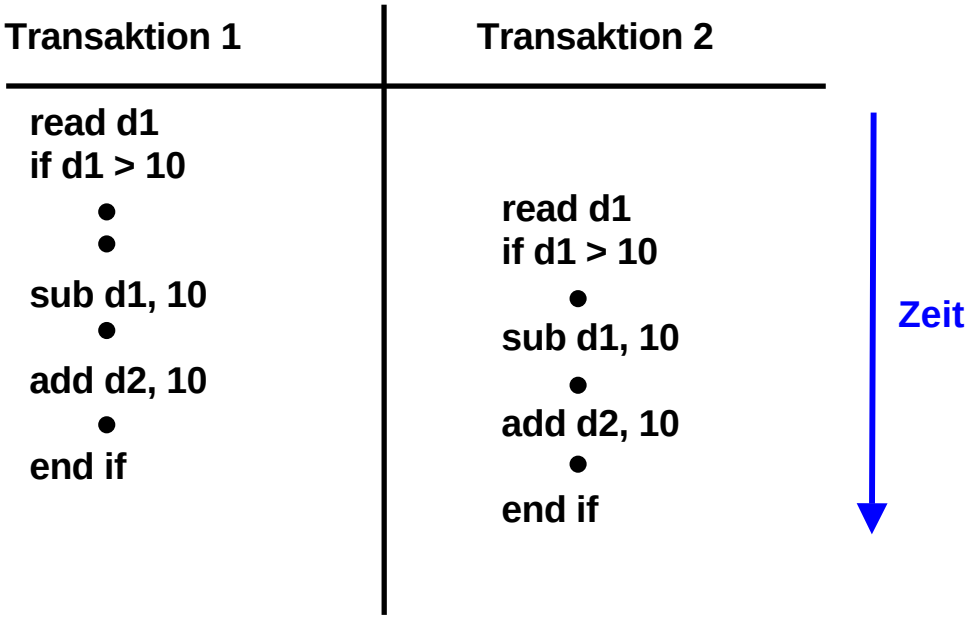
Eine Untermenge der In der Coupling Facility gespeicherten Information ist in den Hauptspeichern der beteiligten Systeme in der Form von Bit Vektoren dupliziert.

Locking Problem

Angenommen zwei Transaktionen, die auf unterschiedlichen Systemen des Sysplex laufen.

Beispiel: Beide Transaktionen greifen auf die zwei Variablen d1 und d2 zu.

Anfangswerte: d1 = 15, d2 = 20 .



Die beiden Abhängigkeiten:

Dirty Read
Eine Transaktion erhält veraltete Information

Lost update
Eine Transaktion überschreibt die Änderung einer anderen Transaktion

müssen gesteuert werden.

Das Ergebnis ist: d1 = - 5, d2 = 40 , obwohl die if-Bedingung ein negatives Ergebnis verhindern sollte.

Concurrency Control

- **Concurrency Control (Synchronisierung) ist zwischen den beiden Transaktionen erforderlich, um dies zu verhindern.**
- **Erforderlich für Application Server, die Daten selbst speichern und auch auf Datenbanken zugreifen.**
- **Der Begriff, der dies beschreibt ist Serialisierung (serializability)**
- **Viele Implementierungsmöglichkeiten. Gebräuchlich ist „two-phase locking“ (nicht zu verwechseln mit dem two-phase Commit Protokoll)**

Two-Phase Locking
Two-Phase Transaktion

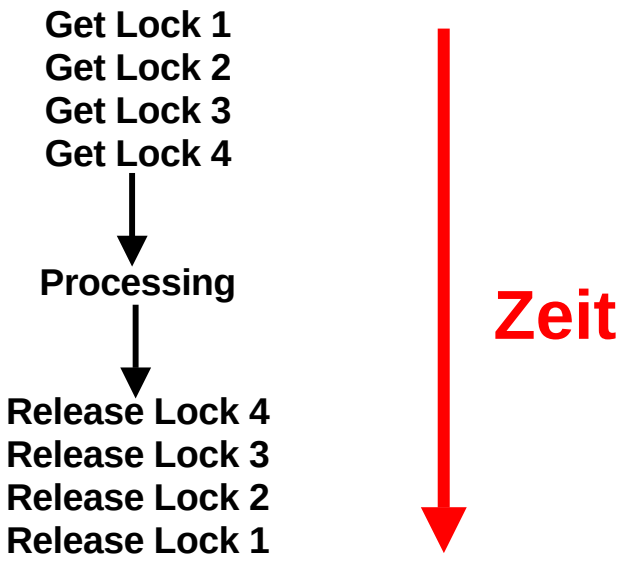
In Transaktionssystemen und Datenbanksystemen werden Locks (Sperren) benutzt, um Datenbereiche vor einem unautorisierten Zugriff zu schützen. Jedem zu schützenden Datenbereich ist ein Lock fest zugeordnet. Ein Lock ist ein Objekt welches über 4 Methoden und zwei Zustände S und E verfügt. Die Methode

- **GetReadLock** **reserviert S Lock (shared),**
- **GetWriteLock** **reserviert E Lock (exclusive),**
- **PromoteReadtoWrite** **bewirkt Zustandswechsel S → E,**
- **Unlock** **gibt Lock frei.**

Mehrere Transaktionen können ein S Lock für den gleichen Datenbereich (z.B. einen Datensatz) besitzen. Nur eine Transaktion kann ein E Lock für einen gegebenen Datenbereich besitzen. Wenn eine Transaktion ein S Lock in ein E Lock umwandelt, müssen alle anderen Besitzer des gleichen S Locks benachrichtigt werden.

Normalerweise besitzt eine Transaktion mehrere Locks.

In einer Two-Phase Locking Transaktion finden alle Lock Aktionen zeitlich vor allen Unlock Aktionen statt. Eine Two-Phase Transaktion hat eine Wachstumsphase (growing), während der die Locks angefordert werden, und eine Schrumpf (shrink) Phase, in der die Locks wieder freigegeben werden.



2-Phase Locking

Es werden alle Locks gesetzt, ehe die Verarbeitung beginnt. Wenn das nicht möglich ist, werden alle Locks sofort wieder freigegeben, und ein neuer Versuch beginnt. Nach Abschluss der Verarbeitung werden die Locks wieder frei gegeben.

Locking Protokoll

Vorgehensweise:

- **Shared Lock (S) erwerben vor dem erstmaligen Lesen**
- **Exclusive Lock (E) erwerben vor dem erstmaligen Schreiben**

derzeitiger Status / Anforderung	kein	Lesen shared	Schreiben exclusive
Lesen Shared	bewilligt, share-mode	bewilligt, share-mode	abgelehnt, Mitteilung über Besitzer
Schreiben Exclusive	bewilligt, exclusive mode	bewilligt, Warnung über Besitzer	abgelehnt, Mitteilung über Besitzer

Mehrere Transaktionen können das gleiche Lock im Zustand S besitzen. Nur eine Transaktion kann ein Lock im Zustand E besitzen.

Eine Transaktion kann ein Lock vom Zustand S in den Zustand E überführen. Hierzu ist es erforderlich, dass eine Nachricht an alle anderen Transaktionen geschickt wird, die das gleiche Lock im Zustand S besitzen. Mittels dieser Nachricht wird das S Lock für ungültig erklärt.

Das Senden einer Nachricht informiert Transaktion 2, dass ihr Wissen über den Zustand der beiden Variablen d1 und d2 nicht mehr gültig ist.

Transaktion 2 muss sich neu über den Zustand der Variablen d1 informieren, ehe sie ein Update von d1 vornimmt.

Frage: Woher weiß Transaktion 1, dass Transaktion 2 ein Interesse an der Variablen d1 hat (ein shared Lock besitzt) ?