UNIVERSITÄT LEIPZIG

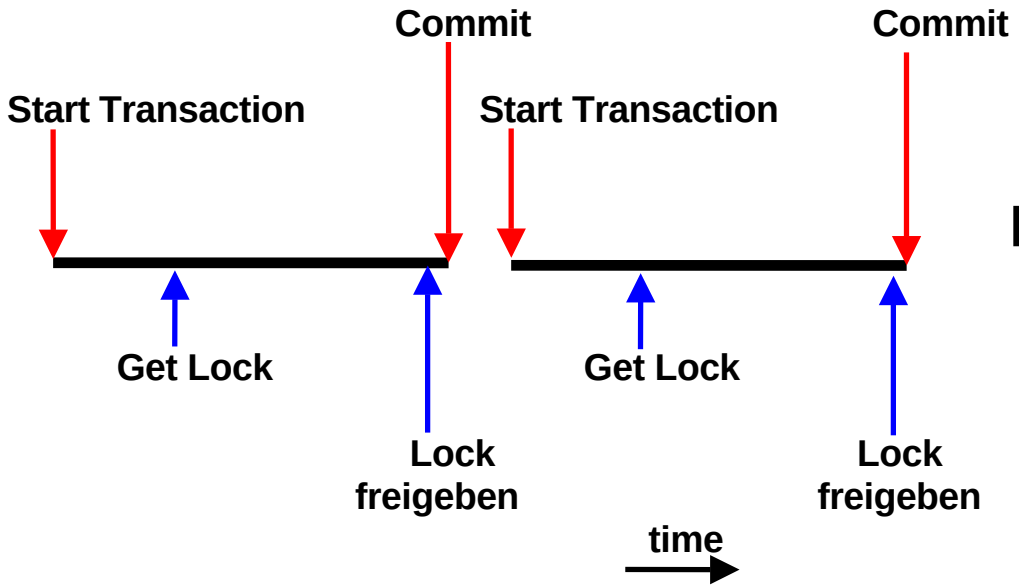
Mainframe Internet Integration

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

SS2013

Parallel Sysplex Teil 4

Cache und Listen Strukturen



Eager Protokoll

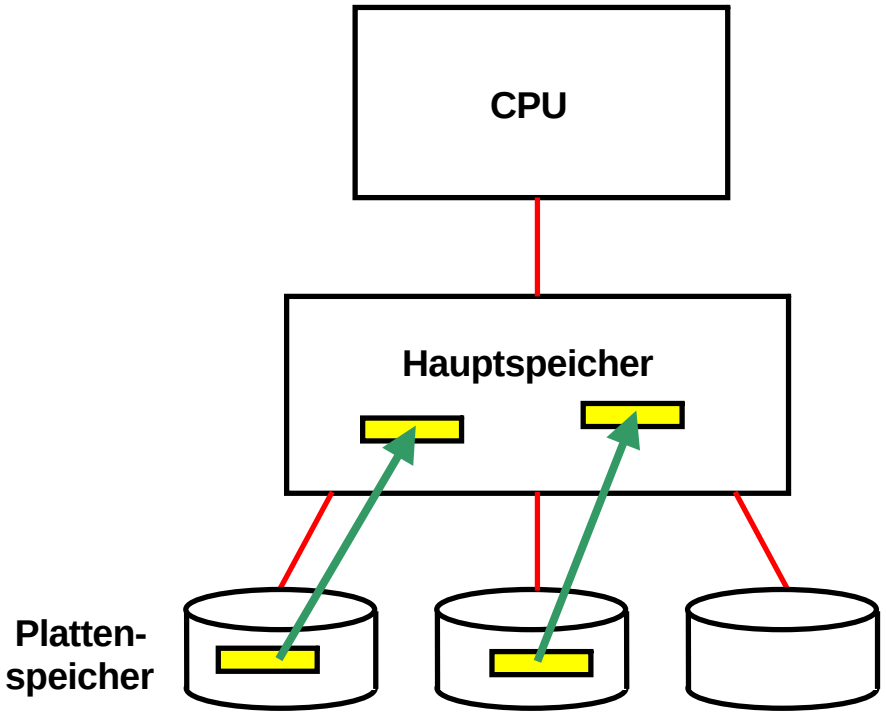
„Eager“ und „Lazy“ Locking Protokolle

Eager Protokoll Lock freigeben wenn Commit die Transaction beendet (siehe das obige Beispiel).

Lazy Protokoll Lock freigeben wenn eine Contention auftritt; nichts tun, bis jemand anderes das Lock benötigt.

Das Lazy Protokoll arbeitet besser, wenn Datenkonflikte selten auftreten. Ein Beispiel ist das TPC-C Benchmark des Transaction Processing Councils (www.tpc.org). Hersteller von Datenbank Software benutzen TPC-C gerne, weil dann ihr Produkt besser aussieht. Die Unterschiede zu praktischen Anwendungen können allerdings sehr groß sein.

Die Sysplex Coupling Facility verwendet das Eager Protokoll (auch als „force-at-commit“ bezeichnet). Datenkonflikte treten häufig auf, wenn existierende Anwendungen auf den Sysplex portiert werden.

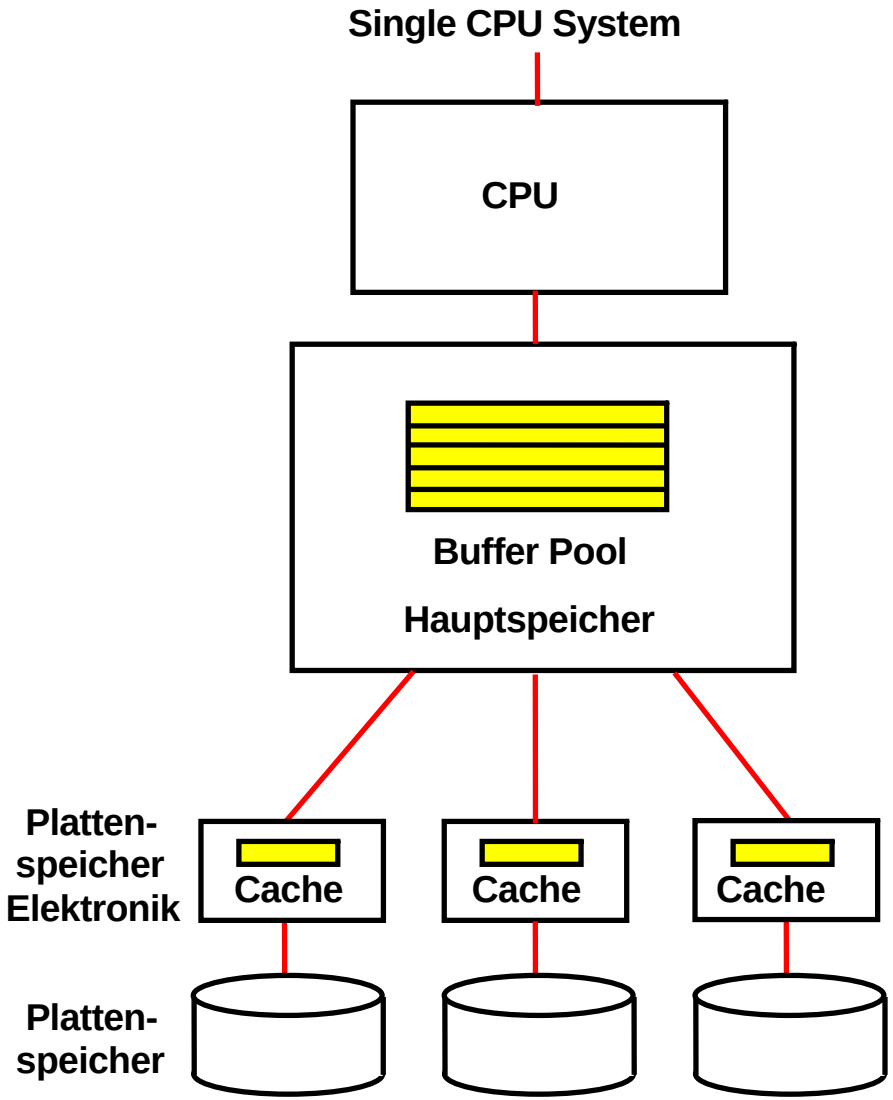


Ein/Ausgabepuffer
I/O Buffer

In der Vergangenheit hat ein Anwendungsprogramm jeweils einzelne Datensätze (Records) vom Plattenspeicher in einen Ein/Ausgabepuffer (I/O Buffer) im Hauptspeicher gelesen und dort verarbeitet. Zur Leistungssteigerung hat man bald mehrere logische Records zu einem physischen Record zusammengefasst. Mit etwas Glück findet das Anwendungsprogramm beim nächsten Zugriff die Daten bereits im Ein/Ausgabepuffer und ein Plattenspeicherzugriff erübrigt sich.

In der Regel wird mit mehreren Dateien oder Datenbanken gleichzeitig gearbeitet, die alle ihren eigenen Ein/Ausgabepuffer benutzen. Die Menge der Ein/Ausgabepuffer wird als „Buffer Pool“ bezeichnet und vom Datenbank-System optimal verwaltet.

"

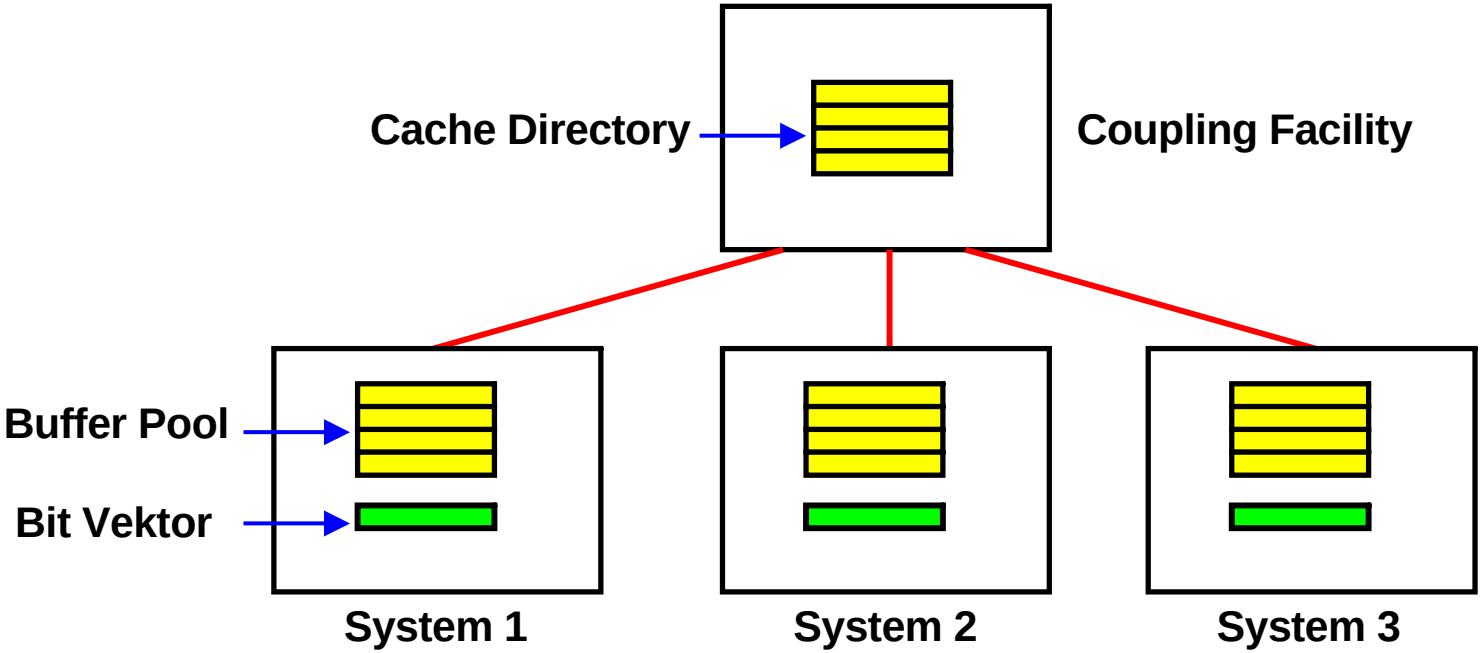


Plattenspeicher Cache und Hauptspeicher Buffer Pool

Der Buffer Pool stellt eine Art Plattenspeicher Cache im Hauptspeicher dar. Das Datenbank-System bemüht sich, den Speicherplatz im Buffer Pool optimal zu verwalten. (Unabhängig davon werden Daten zusätzlich in einem Plattenspeicher Cache gespeichert, der Bestandteil der Plattenspeicher Elektronik oder des Enterprise Storage Servers ist).

Der Buffer Pool besteht aus einzelnen Puffern (Buffers), die Datenbankobjekte oder Teile einer Datei aufnehmen.

In einem Cluster ist nicht auszuschließen, dass Datensätze oder Datenbank Records gleichzeitig in den Buffer Pools mehrerer Knoten (Systeme) abgespeichert werden.



Cache Directory in der Coupling Facility

Der Buffer Pool in jedem System enthält Blöcke (Buffer), die möglicherweise gerade bearbeitet werden. Es kann sein, dass sich in zwei unterschiedlichen Systemen Buffer mit den gleichen Datenbankrecords befinden.

Die Coupling Facility unterhält ein “Cache Directory”, in dem sich jeweils ein Eintrag für jeden Buffer in den angeschlossenen Systemen befinden. Analog zur Lock Verwaltung befinden sich außerdem in jedem System Bit Vektoren, die den Inhalt des Cache Directories teilweise replizieren.

Bei einer Änderung eines Eintrags im Cache Directory erfolgt ein automatisches Update der Bit Vektoren in allen angeschlossenen Systemen.

Coupling Facility Cache Directory

Der lokale Buffer Pool im System 1 enthält Puffer (Blöcke) mit Records, die gerade bearbeitet werden. Solange die Transaktion nicht abgeschlossen ist, verhindert der Lock Manager einen Zugriff durch ein anderes System (z.B. System 2).

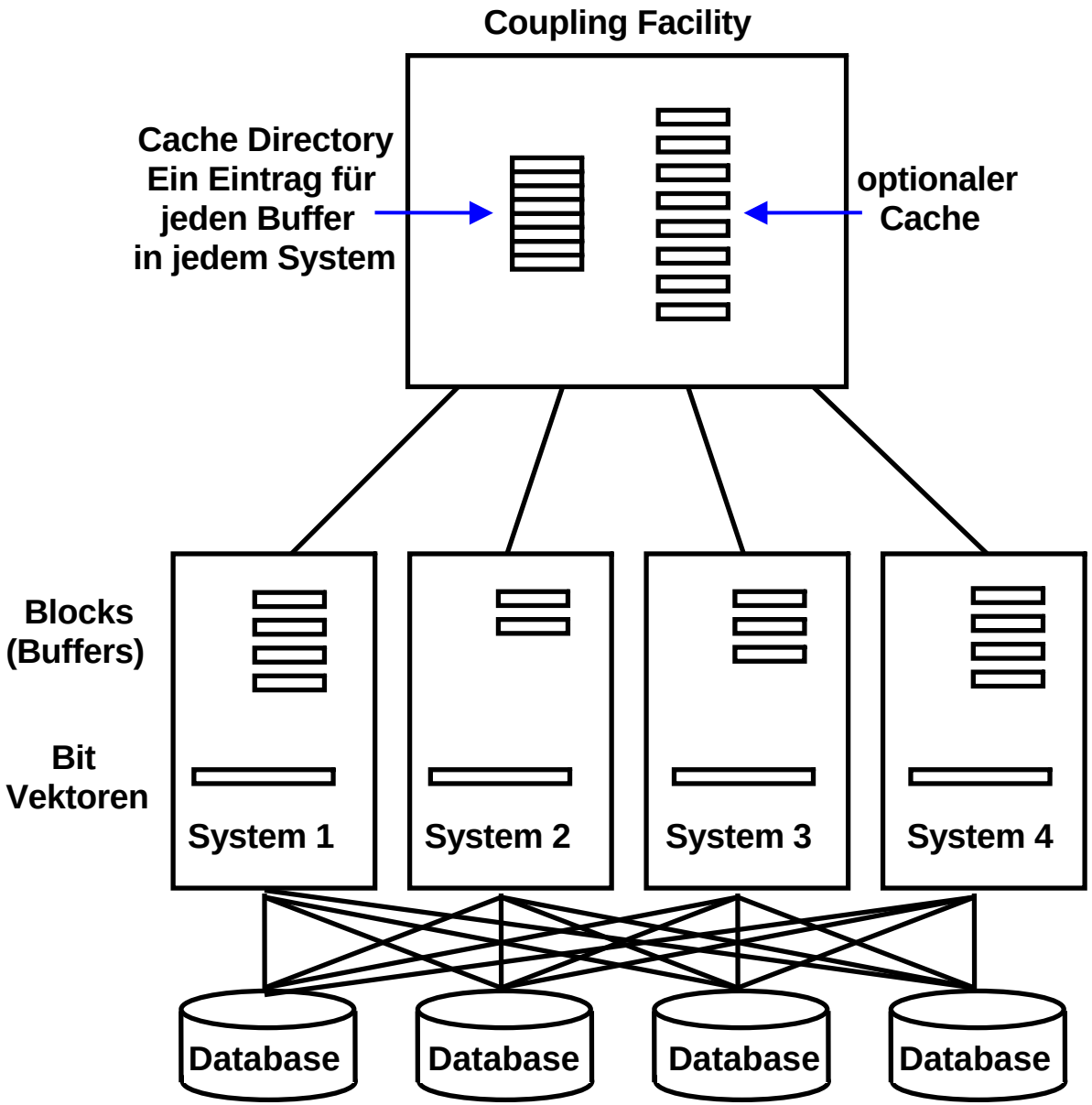
Wenn die Transaktion abgeschlossen ist (commit), werden die Locks freigegeben. Die Puffer bleiben in System 1 erhalten; evtl. werden sie demnächst wieder gebraucht.

Greift System 2 jetzt auf einen Buffer mit dem gleichen Datenbank Record zu, entsteht ein Kohärenzproblem. Die beiden Buffer in den Systemen 1 und 2 haben nicht den gleichen Inhalt.

Lösung: „Force-at-Commit“ . Bei Transaktionsabschluss erfolgt ein update des Cache Directories durch System 1.

Die CF sendet hierzu eine „Cross-Invalidate“ (CI) Nachricht an alle anderen betroffenen Systeme (und nur an die betroffenen Systeme)

Die Cross-Invalidate Nachricht ändert den lokalen State Vector innerhalb des Hauptspeichers eines jeden betroffenen Systems ab. Dies geschieht durch den Link Prozessor und verursacht keine CPU Unterbrechung !



Aller Datentransfer in 4 KByte Blöcken.

Das Cache Directory in der Coupling Facility enthält einen Eintrag für jeden Block (Buffer), der Teil eines Buffer Pools in einem der beteiligten Systeme ist.

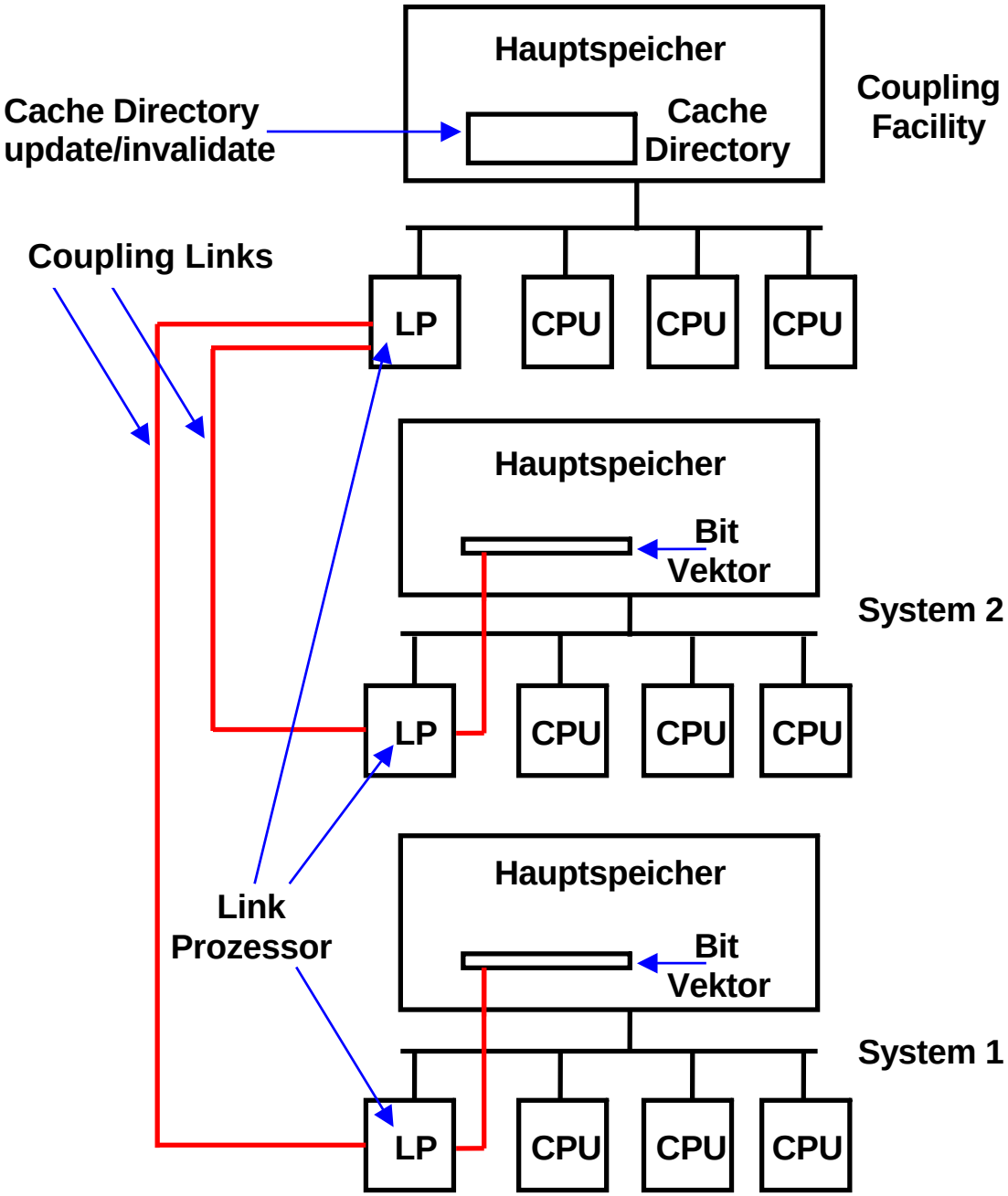
a) System 1 Read from Disk
1. Load Block from Disk
2. Register with CF Directory
3. add Bit in Bit Vector

b) System 2 Read from Disk
1. Load Block from Disk
2. Register with CF Directory
3. add Bit in Bit Vector

einige Zeit später

c) System 1 Intend to write (to local Buffer)
1. Register with CF
2. CF invalidates all Bit Vectors
3. Write to local Buffer

d) System 2 Read from Buffer
1. Read
2. detects invalid Bit im localen Bit Vector
3. führt erforderliche Maßnahmen durch



Beim Force-at-Commit sendet die CF eine „Cross-Invalidate (CI) Nachricht an alle anderen Systeme.

Link Prozessoren (LP) haben einen Direct Memory Access zu dem Hauptspeicher.

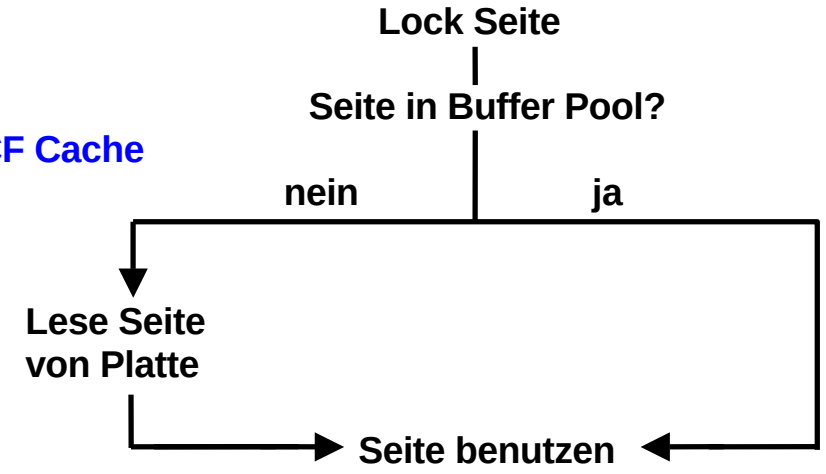
Über die Link Prozessoren der Coupling Facility und die Link Prozessoren der Systeme können Bit Vektoren im Hauptspeicher abgeändert werden, ohne dass der normale Programmablauf in den Systemen dadurch beeinflusst wird (kein Prozesswechsel). Dies verbessert das Leistungsverhalten, da jeder Prozesswechsel eine Pfadlänge von mehreren Tausend Maschinenbefehlen erfordert.

Coupling Facility Cache.

Neben dem Cache Directory kann die Coupling Facility auch als Plattenspeicher-Cache genutzt werden. Hiervon machen einige, aber nicht alle Datenbanksysteme Gebrauch. DB2 und Adabas nutzen die Coupling Facility auch als Plattenspeicher-Cache, IMS jedoch nicht.

DB2, IMS und Adabas sind die wichtigsten unter z/OS eingesetzten Datenbanksysteme.

a) ohne CF Cache

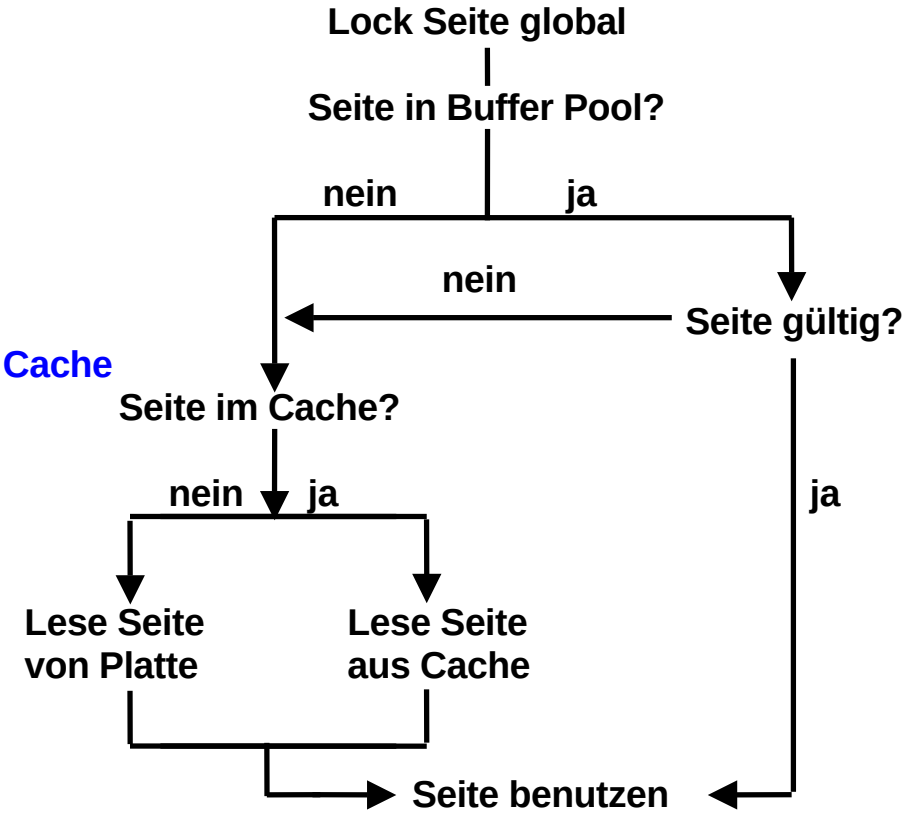


Bei einem Plattenspeicherzugriff werden meistens 4096 Bytes große Blöcke von Daten transportiert (identisch mit der Rahmen Größe).

Ist kein Cache in der Coupling Facility vorhanden, erfolgt der Zugriff auf die folgende Weise:

1. Lock auf die gewünschte Seite setzen.
2. Seite lesen falls im Buffer Pool vorhanden
3. Falls nicht vorhanden, Seite vom Plattenspeicher einlesen.
4. Update des Cache Directories in der CF.

b) mit CF Cache

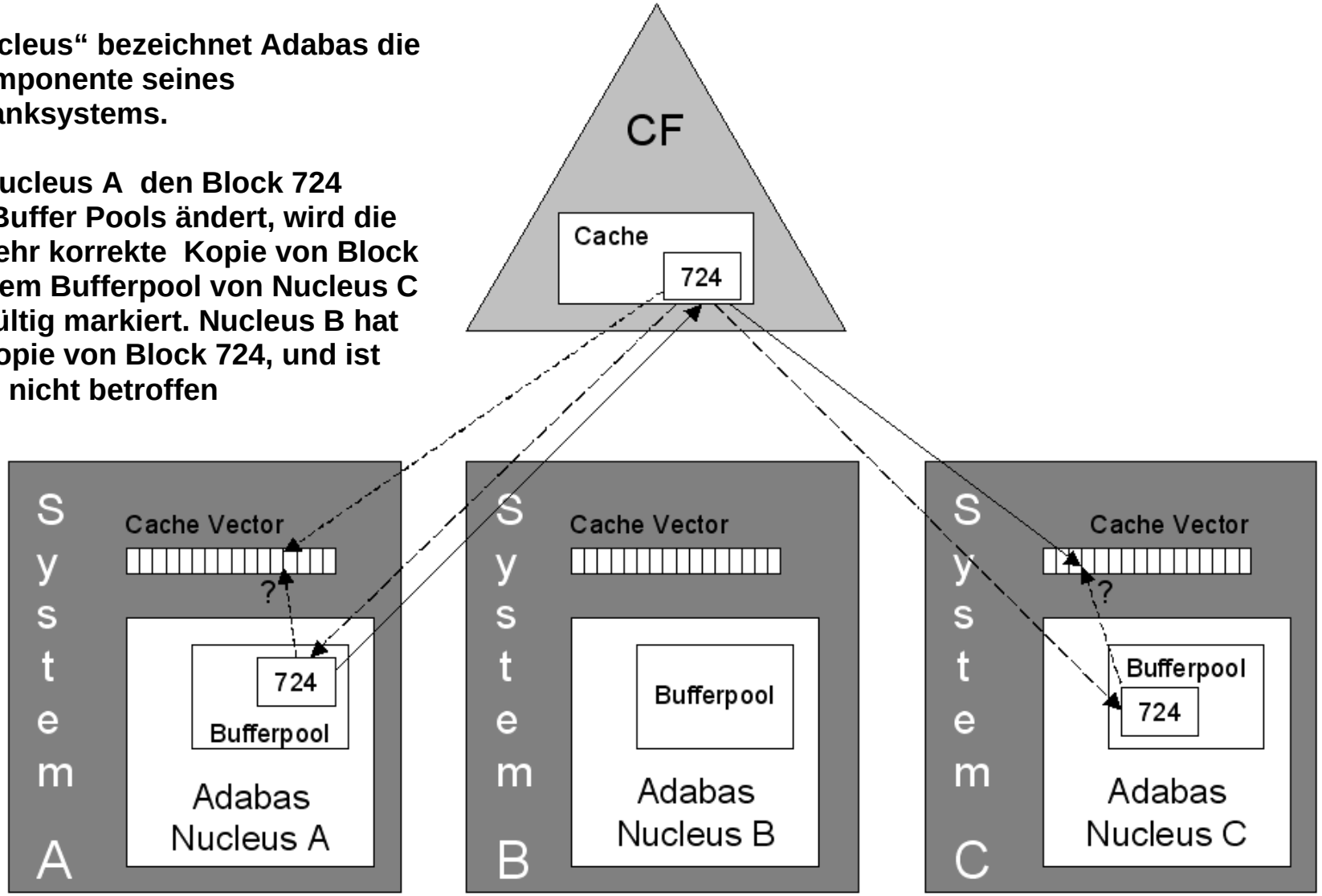


Wird ein Coupling Facility Cache genutzt, wird die Seite aus dem Cache gelesen, falls vorhanden. Wenn nicht, wird die Seite vom Plattenspeicher eingelesen.

In diesem Fall schreibt DB2 zusätzlich die Seite in den CF Cache. Dies ist ein Store-in-Cache; nicht bei jedem Update der Seite wird diese auf dem Plattenspeicher geschrieben. Somit kann die CF Cache Version des Blockes jüngeren Datums sein als die Version auf dem Plattenspeicher.

Als „Nucleus“ bezeichnet Adabas die Kernkomponente seines Datanbanksystems.

Wenn Nucleus A den Block 724 seines Buffer Pools ändert, wird die nicht mehr korrekte Kopie von Block 724 in dem Bufferpool von Nucleus C als ungültig markiert. Nucleus B hat keine Kopie von Block 724, und ist deshalb nicht betroffen



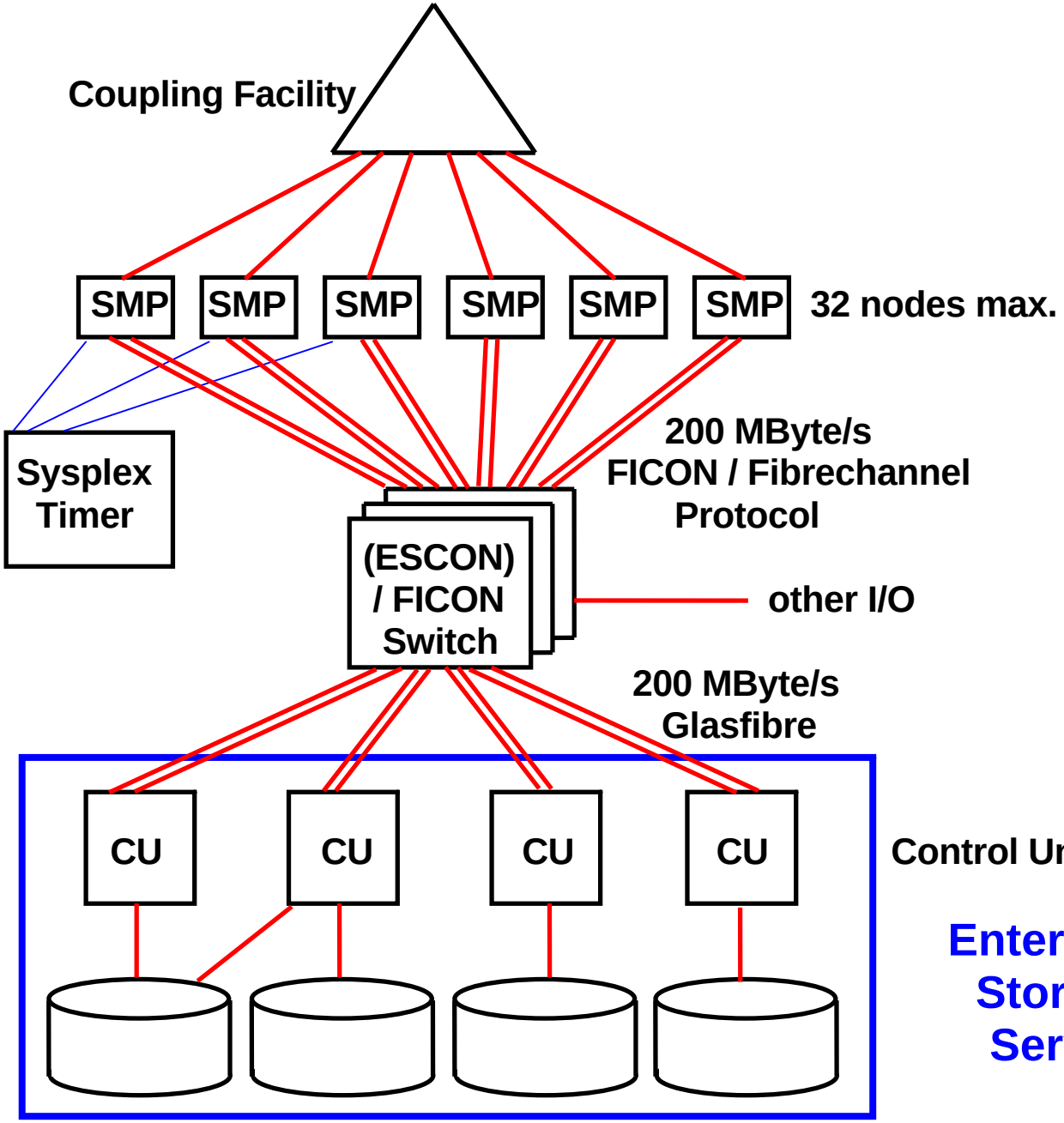
Die Adabas Database der Software AG in Darmstadt nutzt ebenfalls die Coupling Facility als Plattenspeicher Cache.

Adabas

Adabas (Adaptable Database System) wurde von der Software AG, Darmstadt, in 1971 eingeführt. Ursprünglich war ADABAS nur auf Mainframes verfügbar. Heute läuft ADABAS unter den z/OS, VSE, VM/CMS, Fujitsu's Mainframe Betriebssystem Facom, den Fujitsu/Siemens BS2000 Betriebssystemen sowie den Windows, Solaris, AIX, HP-UX, SUSE-Linux und Red Hat Linux Betriebssystemen.

Weltweit wird Adabas von etwa 3000 Installationen benutzt, darunter etwa 100 – 200 z/OS Installationen. Adabas ist ein Geheimtipp für neue Datenbankbenutzer; in Bezug auf Performance nimmt es auch heute noch eine Spitzenposition ein.

Adabas ist kein relationales Datenbanksystem (auch wenn die Firma Adabas behauptet, Adabas sei fast relational). Daten in der ADABAS Datenbank sind hierarchisch in Files strukturiert; eine File entspricht in etwa einer SQL Tabelle. Die Files bestehen aus Records; die Felder der Records entsprechen etwa den Spalten einer SQL Tabelle. Es bestehen einige Ähnlichkeiten mit der IMS Datenbank.



Problem: Die Coupling Facility ist nur mit den Knoten (Systemen) des Sysplex verbunden; sie hat keine direkte Verbindung mit den Plattenspeichern.

Frage: Wenn der Plattenspeicher Cache in der Coupling Facility zu voll wird, wie werden Teile auf einen Plattenspeicher ausgelagert um Platz zu schaffen ?

Cast Out

Wird ein neuer Buffer in die CF Cache geschrieben, muss dafür Platz geschaffen werden und ein anderer CF Puffer auf den Plattenspeicher ausgelagert werden. Die Coupling Facility ist aber nur mit den Systemen (Knoten) verbunden. Sie hat keinen direkten Zugriff auf die Plattenspeicher.

DB2 Instanzen in den einzelnen Systemen unterhalten jeweils einen „Cast-Out“ Thread, der einen Puffer aus dem Cache lesen und auf den Plattenspeicher schreiben kann.

Die Cast-Out Verantwortung wird nach dem Round-Robin Algorithmus (oder einem anderen Algorithmus) den einzelnen Threads zugeordnet. Ein Cast-Out erfolgt jeweils für eine Gruppe von Seiten

CF Cache Recovery

Frage: ist es nicht bedenklich, Daten nur in den Coupling Facility Cache zu schreiben, und nicht sofort auf den Plattenspeicher ?

Ein Duplikat aller Daten befinden sich in den Buffer Pools der einzelnen Systeme. Im Fehlerfall kann hiermit ein Rebuild des CF Cache Inhaltes stattfinden. Weiterhin verfügt eine Installation praktisch immer über 2 CFs, wobei die zweite CF alle Daten der ersten CF dupliziert, und im Fehlerfall automatisch die Funktionen der ersten CF übernehmen kann.

CF List / Queue Strukturen

Neben dem Lock und dem Cache Management enthält die Coupling Facility Listen/Queue Strukturen, die vor allem für eine zentrale Verwaltung aller angeschlossenen Systeme eingesetzt werden.

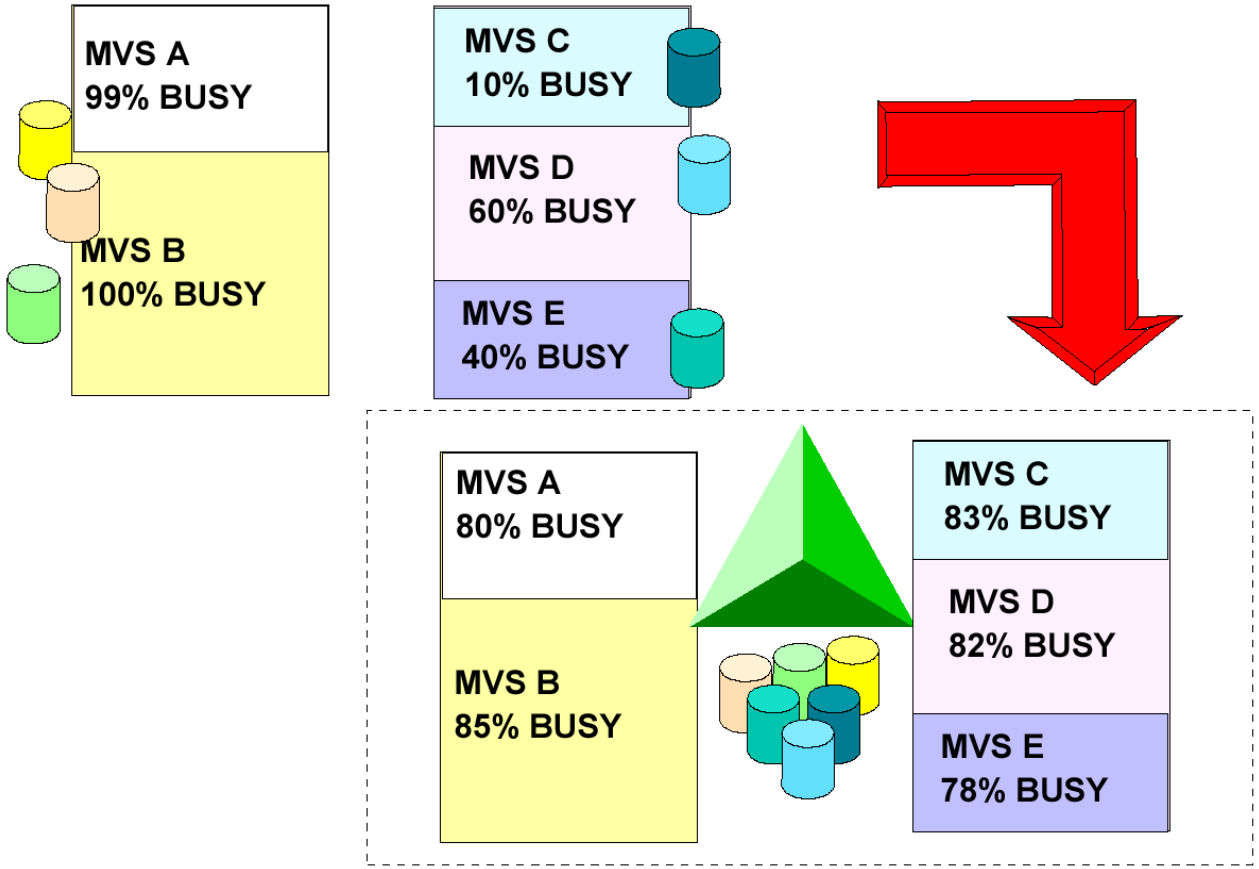
Beispiele hierfür sind:

- **Clusterweite RACF Steuerung. Ein Sysplex Cluster besteht aus mehreren z/OS Instanzen. Im einfachsten Fall müsste sich ein Benutzer mit getrennten Passworten in jedes System einzeln einloggen. „Single Sign On“ ist eine Einrichtung, mit der der Benutzer mit einem einzigen Sign On Zugriffsrechte auf alle Resourcen eines Sysplex erhält. Die entsprechenden RACF Benutzerprofile werden von der Coupling Facility zentral in einer QUEUE/List Struktur verwaltet.**
- **Work Load Management (WLM) Instanzen tauschen periodisch Status Information aus um Transaktionen dynamisch an unterbelastete Systeme weiter zu reichen**
- **MQSeries queue-sharing group eines WebSphere MQ Cluster**

Für einen Zugriff auf die QUEUE/List Strukturen bestehen drei Möglichkeiten:

- **LIFO Queue**
- **FIFO Queue**
- **Key Sequenced**

Key Sequenced bedeutet, dass auf ein bestimmtes Item in einer Queue/List Struktur mittels eines Schlüssels zugegriffen werden kann.

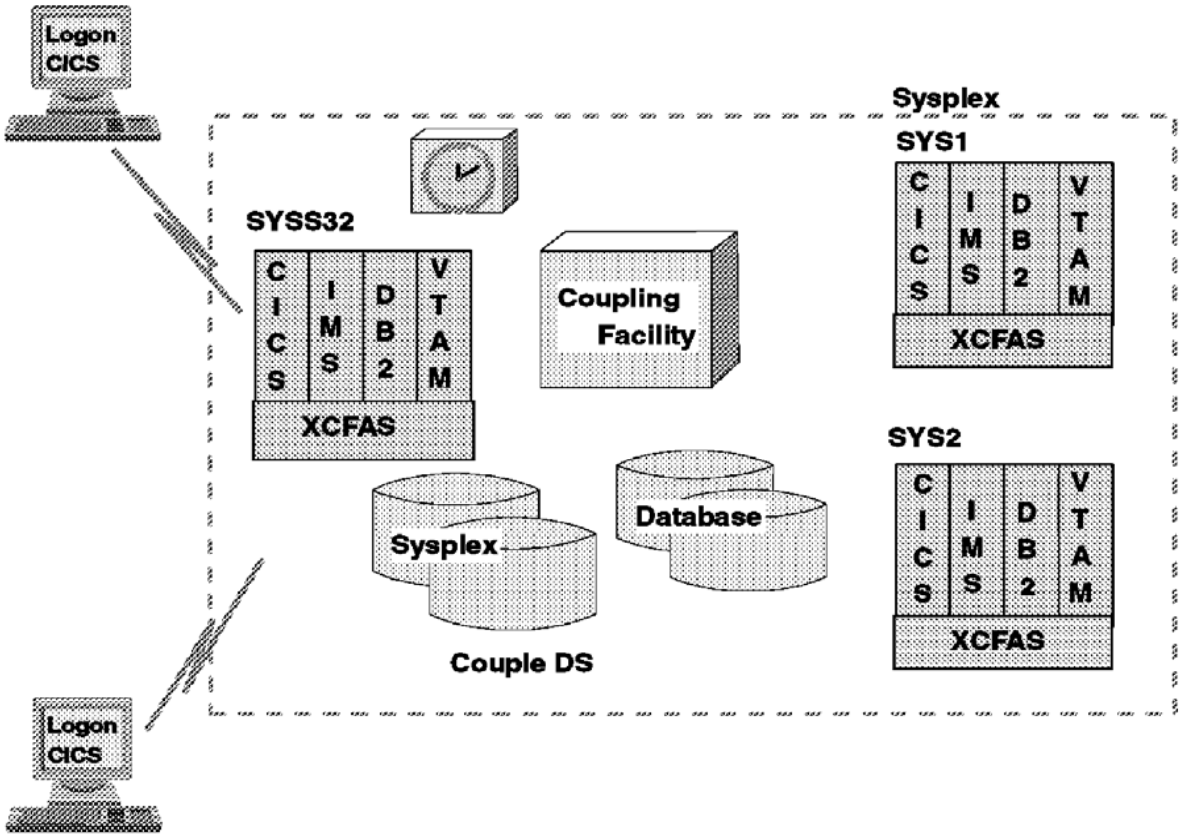


Work Load Balancing

Gezeigt sind 5 Systeme: MVS A , MVS B ,MVS C ,MVS D und MVS E .

Es ist fast unausbleiblich, dass diese Systeme unterschiedlich ausgelastet sind.

Indem man die Rechner zu einem Sysplex zusammenfasst, kann eine Systemkomponente, der „Work Load Manager“ (WLM) für eine gleichmäßige Auslastung aller 5 Systeme sorgen. Die entsprechenden Daten werden in der CF gespeichert.



Cross Coupling Facility Address Space: XCFAS

Angenommen mehrere Instanzen einer Anwendung oder eines Subsystems auf unterschiedlichen Knoten eines Sysplex, z.B. CICS oder WebSphere. Mit Hilfe von XCFAS können die Instanzen Status Information austauschen oder miteinander kommunizieren.

Die gemeinsam genutzten Daten befinden sich als Listen- oder Queue-Strukturen auf der Coupling Facility. Der Zugriff auf diese Daten erfolgt mit Hilfe des Cross-System Extended Services (XES) Protokolls, welches Zugriffs- und Verwaltungsdienste zur Verfügung stellt.

Sysplex Performance

... und wie sieht es mit dem Leistungsverhalten und der Skalierbarkeit eines Sysplex aus ?

In den meisten Fällen ist es sehr schwierig, eine existierende Anwendung, die auf einer einzelnen CPU läuft auf einen Mehrfachrechner zu portieren. In vielen Fällen bringt bei 10 oder 30 CPUs jede weitere CPU keinen nennenswerten Leistungsgewinn mehr.

Der Sysplex und besonders die Coupling Facility bilden hier eine Ausnahme. Sie weisen hervorragende Skalierungseigenschaften auf. Dies wird in den beiden folgenden Abbildungen dargestellt.

Heute (2011) ist die Coupling Facility immer noch ein Mainframe Alleinstellungsmerkmal. Wir erwarten, dass in Zukunft die Geschwindigkeit einer einzelnen CPU nur noch langsam steigen wird. Leistungssteigerungen unser Rechner sind daher in Zukunft in erster Lininie durch den Einsatz von Mehrfachrechnern zu erwarten.

Es ist daher vorstellbar, dass eine Coupling Facility in Zukunft auch für die x86 Plattform verfügbar sein wird – warten wir es ab.

Sysplex Overhead

Installation	Anzahl Systeme	% Sysplex Overhead
A	4	11 %
B	3	10
C	8	9
D	2	7
E	11	10
DB2 Warehouse Workload	2	13

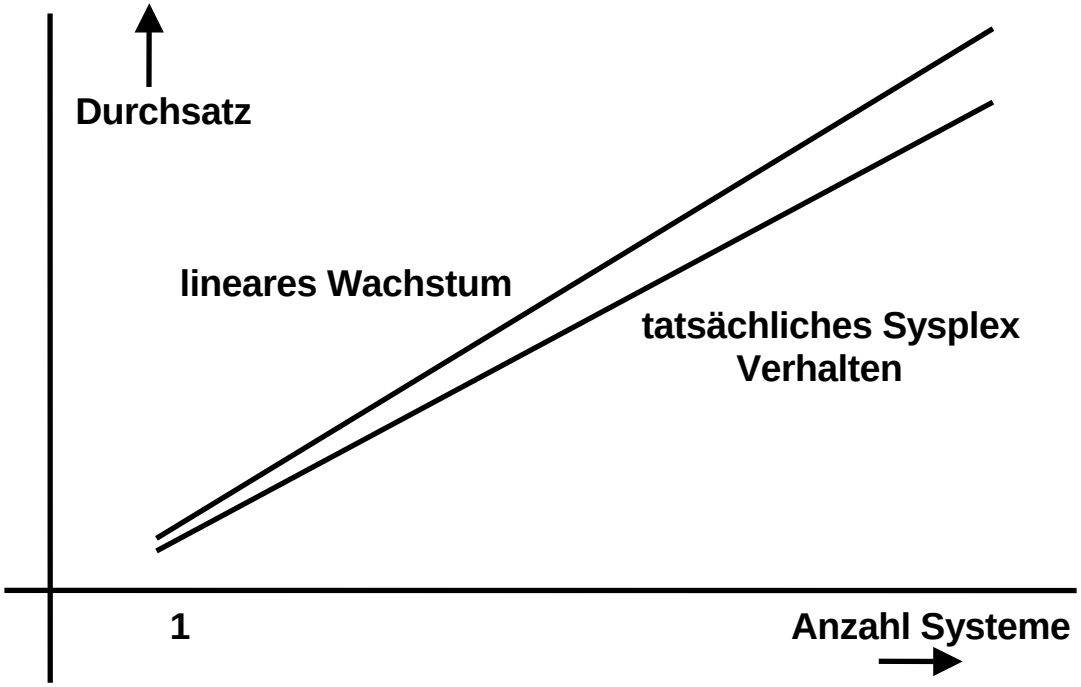
Gezeigt ist der durch die Sysplex Support Software verursachte Overhead (zusätzlich benötigte CPU Zyklen). Die Messungen erfolgten auf den Sysplex Installationen von 5 großen Unternehmen, hier als A, B, C, D und E bezeichnet. Weiterhin erfolgten Messungen auf einer IBM internen DB2 Warehouse Installation.

Der Overhead bewegt sich zwischen 7 und 13 %.

Die Sysplex Support Software (wenn installiert) erzeugt in jedem System zusätzlichen Overhead, selbst wenn der Sysplex nur aus einem einzigen System besteht. In jedem System wird zusätzliche CPU Kapazität benötigt um den gleichen Durchsatz zu erreichen.

Die Sysplex Support Software wird nur dann installiert, wenn der Rechner als Bestandteil eines Sysplex genutzt wird.

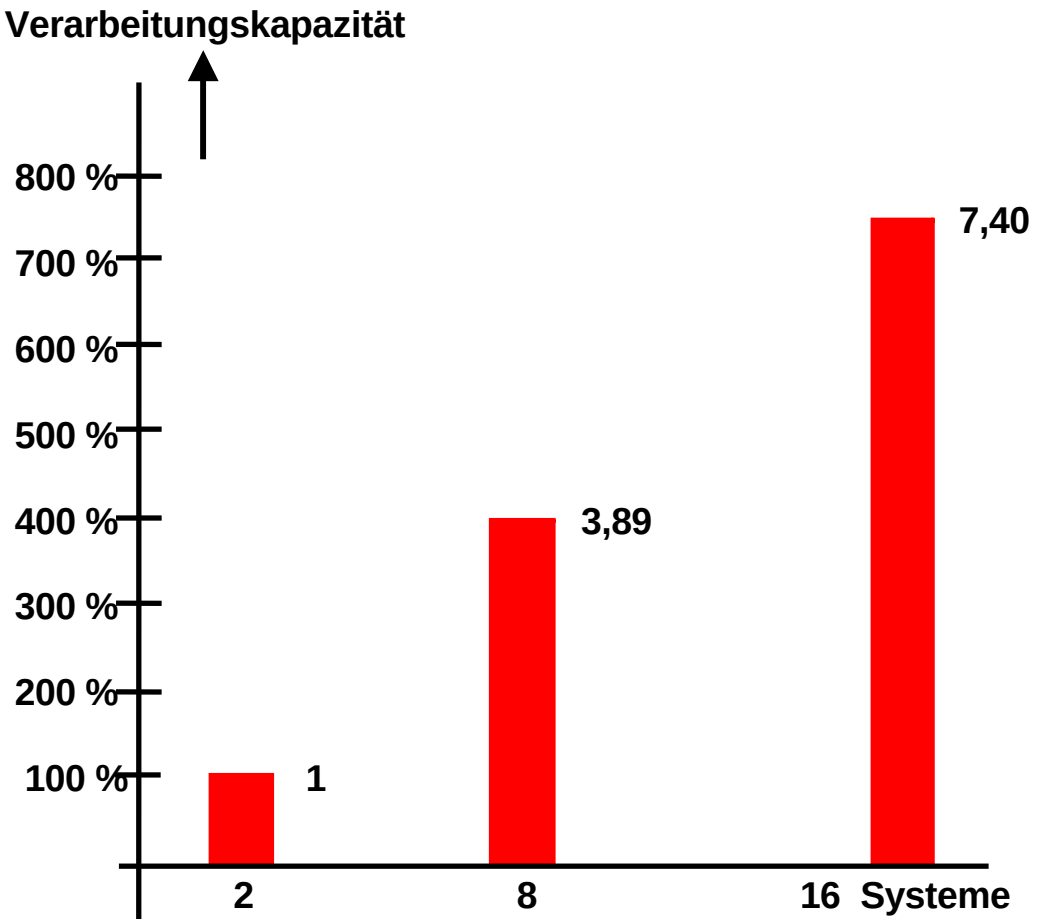
Sysplex Overhead



Mit nur einem einzigen System entsteht der erwähnte Sysplex Support Software Overhead. Mit einer wachsenden Anzahl von Systemen beobachten wir eine lineare Skalierung.

“The Parallel Sysplex environment can scale nearly linearly from 2 to 32 systems. The aggregate capacity of this configuration meets every processing requirement known today.”

http://publib.boulder.ibm.com/infocenter/zos/basics/index.jsp?topic=/com.ibm.zos.zconcepts/zconc_pllsysscale.htm



Mit 2 Systemen, die mit einer CF verbunden sind, erreicht man eine Leistung von 100 %.

Mit 16 Systemen sollte man theoretisch eine Leistung von 800 % erreichen. Tatsächlich erreicht werden 740 %.

Jedes System kann z. B. 8 oder 16 CPUs enthalten.

Dies ist ein sehr guter Skalierungswert.

Parallel Sysplex Leistungsverhalten

Testergebnisse einer Installation bestehend aus CICS Transaktionsmanager, CICSplex System Manager, IMS Datenbank, Mit einer Mischung von OLTP, Reservierung, Data Warehouse und Bankanwendungen.

Literatur: Coupling Facility Performance: A Real World Perspective, IBM Redbook, March 2006,
http://www.redbooks.ibm.com/redpapers/pdfs/redp4014.pdf