

Андрей Криушин
Консультант по ПО Oracle
компании РДТЕХ

ORACLE | CERTIFIED
PROFESSIONAL



Потенциал технологии **Oracle9i Real Application Clusters**

Темы семинара

- ✓ Архитектура
- ✓ Предназначение
- ✓ Особенности реализации
- ✓ Особенности применения

Распараллеливание задач

Функциональный параллелизм – множество независимых задач

Параллелизм данных - однотипные операции над независимыми наборами данных

Внешний параллелизм операций – конвейерная обработка (pipeline)

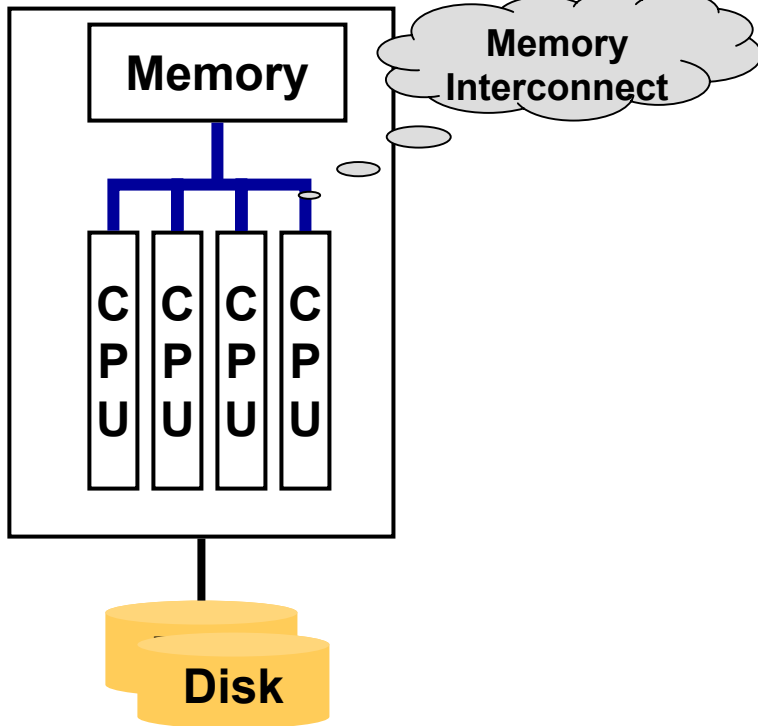
Внутренний параллелизм операций

- ✓ сортировки – могут быть распараллелены за счет разбиения исходного набора данных на подмножества, каждое из которых сортируется независимо, затем результаты «досортируются»
- ✓ ассоциативные операции (суммирование и т.п.) – можно вычислить частные суммы по подмножествам данных, а потом сложить результаты

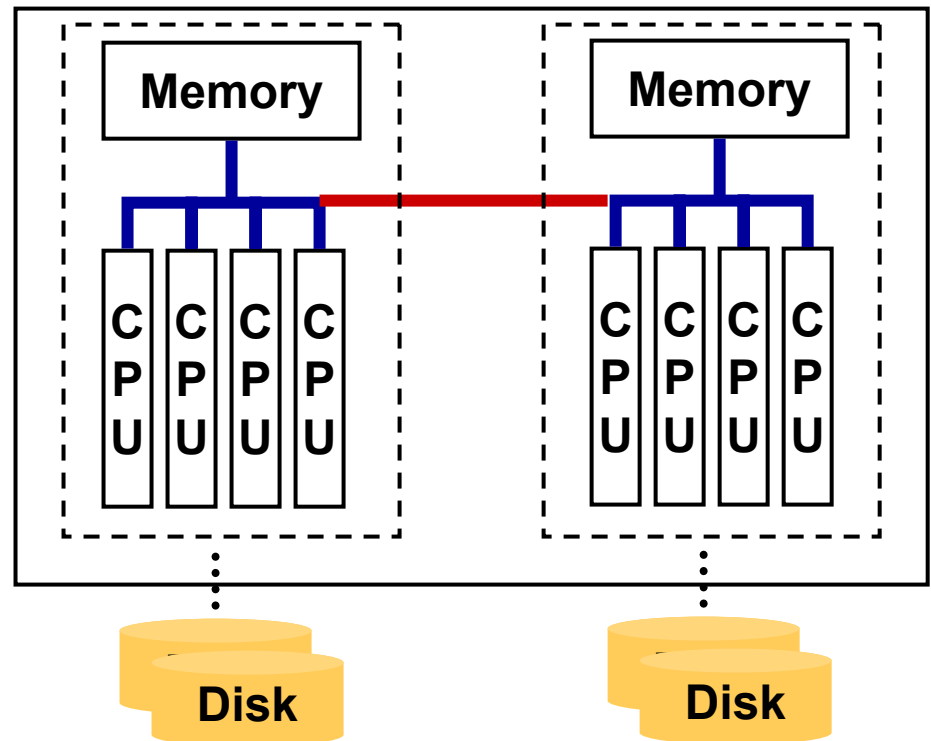
Аппаратные «параллельные» архитектуры

SMP / NUMA

Symmetric
Multi-Processing
(SMP)

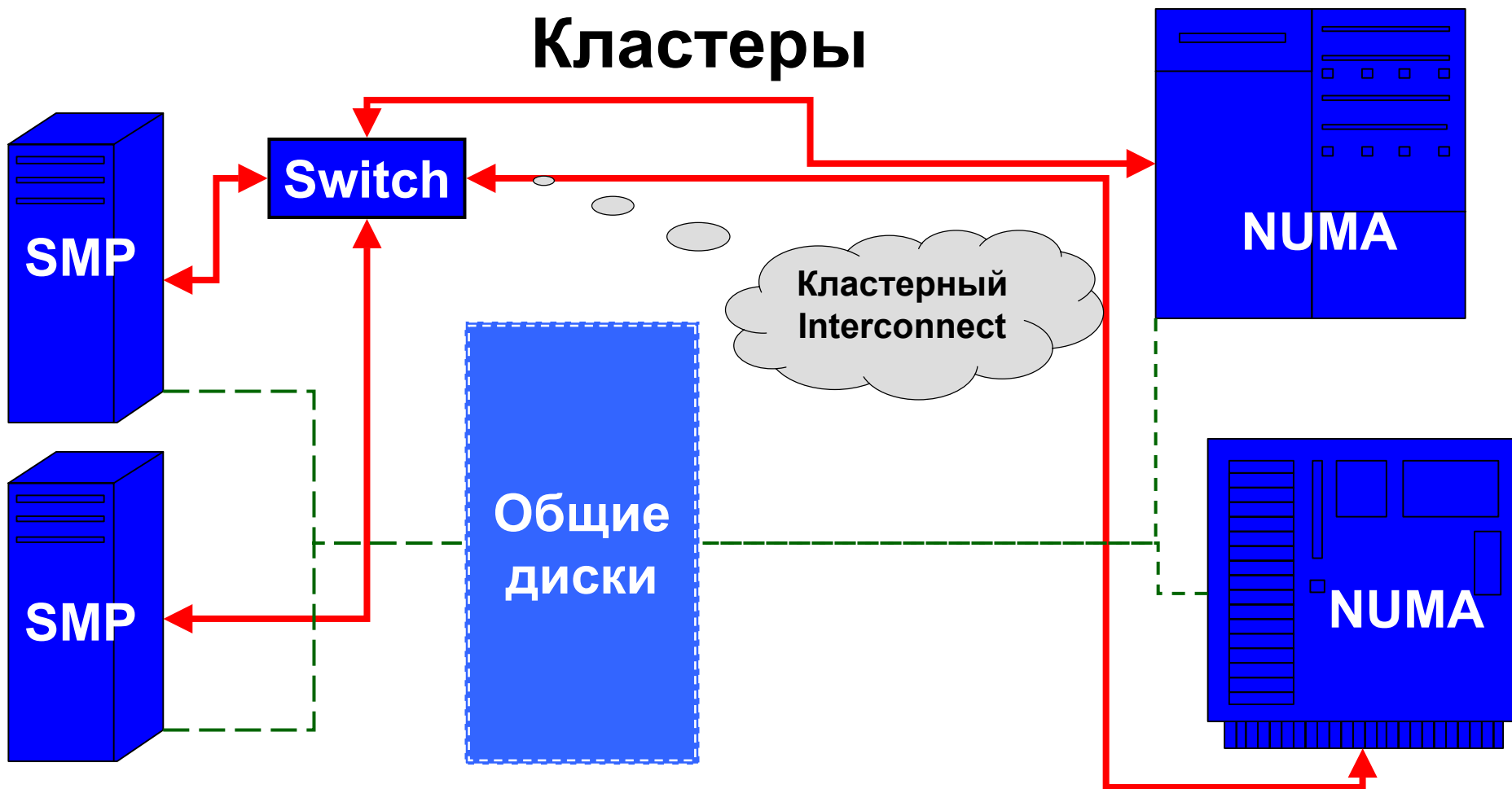


Non-Uniform
Memory Access



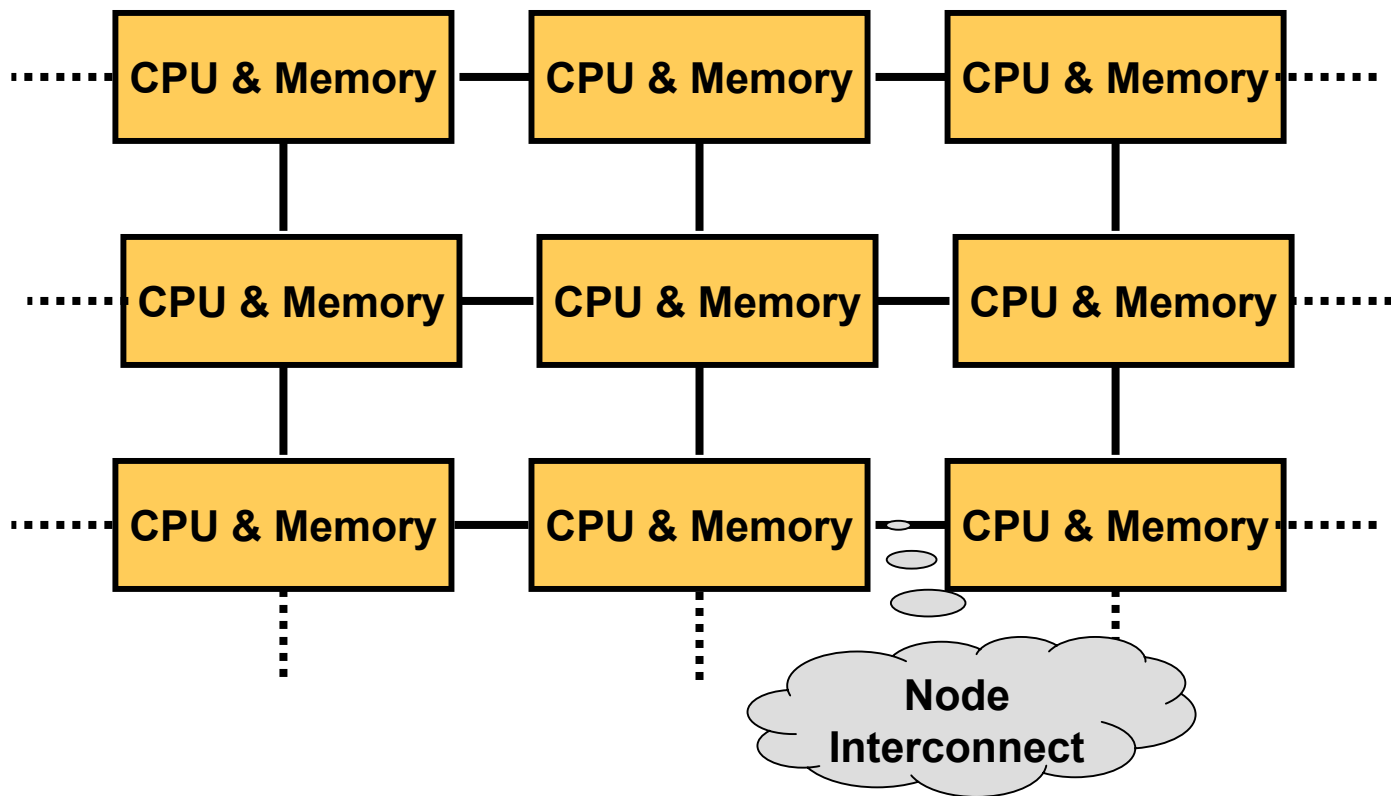
Аппаратные «параллельные» архитектуры

Кластеры



Аппаратные «параллельные» архитектуры

Massively Parallel Processing (MPP)



Аппаратная архитектура кластера для Oracle9i RAC



Программная архитектура ОС для Oracle9i RAC

OSD – Operating System Dependent Layer

- ✓ **Cluster Manager (CM)** – поставляется Oracle или производителем системного ПО
 - ✓ **Node Monitor** (компонент CM)
 - ✓ **IPC (Inter Process Communication)**
 - Примеры: TCP/IP через 1Gbit/sec Ethernet, Sun Fire™ Link с соответствующим ПО
 - ✓ **Разделяемый доступ к дискам**
 - Драйверы ОС для работы с SAN/NAS/SCSI/FireWire, диспетчер логических томов (Logical Volume Manager)
 - Информация о структурах хранения на разделяемых дисках (тома, кластерные файловые системы и так далее) должна быть доступна всем узлам кластера
- База данных может быть создана на**
- неформатированных разделах/томах (raw devices)
 - кластерной файловой системе (если доступна для данной платформы)
- Кластерные файловые системы**
- OCFS (Oracle Cluster File System) для MS Windows и Linux (<http://otn.oracle.com>)
 - GFS (Global File Services) для Sun – только для ПО Oracle и архивных журналов

Не путать с протоколом Oracle Net, имеющим то же название

Программная архитектура СУБД Oracle для Oracle9i RAC

- ✓ Два и более экземпляра работают с одной базой данных
- ✓ Необходима координация действий экземпляров
 - ✓ «Глобальная служба кеша» («Global Cache Services», **GCS**) обеспечивает скоординированный доступ к блокам БД в кешах буферов разных экземпляров
 - ✓ «Глобальная служба блокировок» («Global Enqueue Services», **GES**) управляет транзакционными блокировками (на уровне словаря данных) и другими общими ресурсами (например, SCN)
- ✓ **GCS и GES**
 - ✓ реализованы в виде дополнительных фоновых процессов экземпляра (LMSn, LMON, LMD, LCKn, DIAG)
 - ✓ работают с распределенной по экземплярам базой «глобальных ресурсов» посредством «глобальных блокировок»
 - ✓ В предыдущих версиях – DLM (Distributed Lock Manager)
- ✓ **GCS использует механизм «Cache Fusion»** для обмена содержимым кеша буферов между экземплярами через высокоскоростной Interconnect, что существенно расширяет возможности масштабирования приложений, которые невозможно было использовать в среде Oracle Parallel Server в предыдущих версиях ПО Oracle

Основные задачи, решаемые Oracle9i RAC

✓ Масштабируемость

- Распределение нагрузки по узлам кластера
- Возможность постепенного наращивания производительности системы по мере роста потребностей в ресурсах за счет добавления узлов кластера

✓ Отказоустойчивость

- Сбой не приводит к полной недоступности системы, поскольку предполагается, что некоторое количество узлов/экземпляров не будет затронуто сбоем
- Время недоступности в случае сбоя – наименьшее из всех поставляемых корпорацией Oracle решений по обеспечению высокой доступности

✓ Прозрачность

- С точки зрения пользователей и приложений, база данных в среде Oracle9i RAC выглядит как обычная БД под управлением одного экземпляра
- ПО для администрирования (Oracle Enterprise Manager) и резервирования & восстановления (Recovery Manager) адаптированы к среде Oracle9i RAC
- Администрирование БД имеет дополнительные особенности. Важно, что БД – одна (сравните с распределенными базами данных, репликацией и так далее)
- Рутинные операции (добавление пользователей, табличных пространств) не требуют дублирования на других экземплярах.

Масштабируемость в предыдущих версиях Oracle (Oracle Parallel Server, OPS)

✓ Oracle 7.3 – 8.0.6

- Наибольшие проблемы связаны с синхронизацией кешей буферов
- **Полностью масштабируемо** только одновременное чтение (**SELECT**)
- **Синхронизация модифицированного буфера** осуществляется путем его принудительной записи на диск одним экземпляром и последующим чтением блока с диска другим экземпляром
- Перенос OLTP в среду OPS был рискованным занятием, и, как правило, требовал адаптации приложения, например, добавления «служебных» столбцов в таблицы, индексы и прочее

✓ Oracle8i

- По-прежнему, наибольшие проблемы связаны с синхронизацией кешей буферов
- Добавлен механизм «Cache Fusion Phase I» – если блок модифицирован в кеше одного экземпляра, но требуется другому экземпляру для чтения (SELECT), то первый экземпляр конструирует образ блока для чтения и пересылает его второму через Interconnect, минуя диск
- Появилась возможность разделения экземпляров по типу активности, то есть один экземпляр несет нагрузку OLTP, а другой – DSS/OLAP

Масштабируемость в Oracle9i RAC для DSS/OLAP

DSS – Decision Support Systems, «отчеты»

OLAP – On-Line Analytical Processing, «оперативный анализ»

Минимум модификаций, в основном ресурсоемкие SELECT'ы

РЕЗУЛЬТАТ: уменьшение времени отклика (speedup)

- ✓ Наилучший кандидат для переноса в среду Oracle9i RAC
- ✓ Ускорение достигается за счет распараллеливания отдельной команды SQL по нескольким процессам на *разных* экземплярах RAC (в *дополнение* к распараллеливанию в рамках одного экземпляра)
- ✓ Настраивается автоматически
- ✓ Можно определять группы экземпляров, участвующих в распараллеливании
- ✓ Без необходимости не производится (instance affinity, привязка к экземпляру)
- ✓ Дополнительные ресурсы CPU и памяти

Масштабируемость в Oracle9i RAC для OLTP

OLTP – On-Line Transaction Processing, множество коротких транзакций, в основном, модификации базы данных

РЕЗУЛЬТАТ: увеличение количества одновременно выполняемых транзакций (scale-up)

- ✓ Больше процессов ОС – больше сеансов БД
- ✓ Больше потоков журнальной информации – меньше ожиданий записи в журнал
- ✓ Дополнительные ресурсы CPU и памяти

Возможны проблемы масштабируемости

- ✓ если сеансы разных экземпляров часто модифицируют одни и те же блоки (Cache Fusion,

Масштабируемость в Oracle9i RAC для OLTP

Проблемы плохо спроектированного приложения
НЕ РЕШАЮТСЯ переходом в среду Oracle9i RAC.
Если на одном экземпляре производительность
ограничивалась ожиданиями «buffer busy waits»,
то в Oracle9i RAC будет еще хуже!!!

Масштабируемость гибридных систем (смесь OLTP и DSS/OLAP)

- ✓ **Справедливо** всё вышесказанное о OLTP и DSS/OLAP по отдельности
- ✓ **Интересная особенность** состоит в том, что если на уровне приложения разделить OLTP и DSS/OLAP по разным узлам кластера, то выигрыш могут получить и OLTP, и DSS/OLAP
- ✓ **Пример** — если выяснится, что подсистема OLTP масштабируется плохо, можно оставить OLTP на одном экземпляре, но разгрузить его за счет выноса DSS/OLAP на отдельный узел или узлы
- ✓ **Очень важно**, что параметры экземпляра, такие как размер кеша буферов, разделяемого пула и так далее, могут настраиваться независимо для разных экземпляров. В обычном сервере приходится искать *компромисс* между требованиями OLTP и DSS/OLAP

Решения Oracle для систем высокой доступности (High Availability)

- ✓ Oracle9i Real Application Clusters
- ✓ Advanced Replication Option
- ✓ Cluster Guard (Oracle FailSafe) или его аналоги
- ✓ Резервная база данных
(Oracle DataGuard, STANDBY)

Отказоустойчивость Oracle9i RAC

Oracle9i RAC не защищает от сбоя носителя!

Требуется избыточность аппаратных компонент

Особенно для дисковой подсистемы

- ✓ Восстановление незавершенных на «сбойном» экземпляре транзакций производится уцелевшими экземплярами автоматически
- ✓ Сбой не влияет на пользователей, подсоединенные к уцелевшим экземплярам
- ✓ Время недоступности системы для пользователей, подсоединенных к сбойному экземпляру, равно времени переключения на один из уцелевших экземпляров

Отказоустойчивость Oracle9i RAC

- ✓ **T**ransparent **A**pplication **F**ailover (с Oracle 8.0)
- ✓ Standby instance - для кластера из двух узлов есть возможность сконфигурировать один из экземпляров как «резервный»
 - соединение по Oracle Net невозможно, только локальные подключения
 - задания (DBMS_JOBS) работают
 - можно «разогреть» кеш команд (пакет DBMS_LIBCACHE)
 - «разогрев» кеша команд позволяет избежать замедления работы, характерного для ситуации сразу после перезапуска экземпляра

Отказоустойчивость Oracle9i RAC

Сравнение с обычным сервером и Cluster Guard

- ✓ **Сбой экземпляра** требует перезапуска на том же или другом узле кластера (Cluster Guard)

Время недоступности системы для всех пользователей

- Время запуска экземпляра
- Время наката оперативных журналов
- После перезапуска наблюдается замедление работы, поскольку кэши команд и буферов пусты

Типичное время восстановления – минуты

- ✓ **Сбой ОС требует перезагрузки сервера**

Для Sun 15K это около часа

- ✓ **НЕДОСТАТОК:** сервер (идентичный основному) занимается исключительно ожиданием сбоя основного сервера, то есть его ресурс недоступен для текущей активности базы данных

Отказоустойчивость Oracle9i RAC

Сравнение с DataGuard (STANDBY)

Защищает от сбоя носителя!

- ✓ **Требуется передача всей журнальной информации**
- ✓ **В Oracle9i** можно сконфигурировать Oracle DataGuard в режимах:
 - MAXIMUM PROTECTION (нулевая потеря зафиксированных транзакций)
 - MAXIMUM AVAILABILITY (почти нулевая потеря, за исключением времени недоступности STANDBY)
 - MAXIMUM PERFORMANCE (журнальная информация передается по мере заполнения журнальных файлов основной БД)Резервная БД может быть открыта в режиме **READ ONLY**. В это время её восстановление не производится
- ✓ **Переключение на STANDBY** требует **принятия решения**
- ✓ **НЕДОСТАТОК:** сервер (идентичный основному) занимается исключительно восстановлением STANDBY, то есть его ресурс недоступен для текущей активности БД
- ✓ **Время недоступности** для всех пользователей равно:
 - a) время на принятие решения,
 - b) активация STANDBY
 - c) переключение сетевых компонент
- ✓ **Типичное время восстановления** – до десятка минут

«Глобальные ресурсы» и «Глобальные блокировки»

- ✓ **«Глобальные ресурсы»** – каждому элементу экземпляра, требующему согласованного доступа, назначается «глобальный ресурс»
 - ✓ Блок базы данных
 - ✓ SCN
 - ✓ Последовательность (Sequence)

- ✓ **«Глобальные блокировки»**

- Доступ к любому «глобальному ресурсу» может быть выдан экземпляру в одном из режимов блокировки – eXclusive (X), Shared (S), Null (N)
- Не все режимы совместимы

Режимы	X	S	N
X	-	-	+
S	-	+	+
N	+	+	+

- Если экземпляру требуется ресурс в несовместимом режиме, экземпляр затребует от других экземпляров (через LMD – диспетчера «глобального ресурса») преобразования «глобальной блокировки» в совместимый режим
- Преобразования «глобальных блокировок» требуют обмена сообщениями и определенных действий со стороны других экземпляров

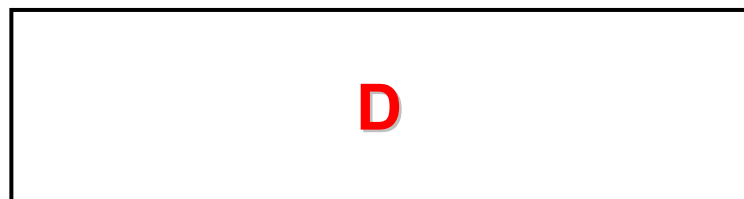
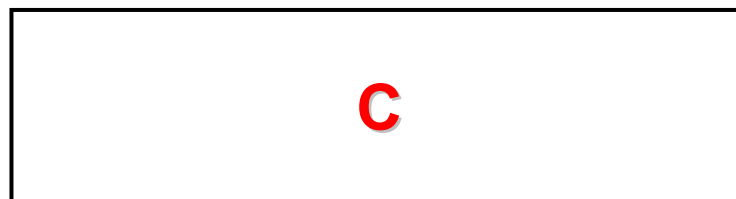
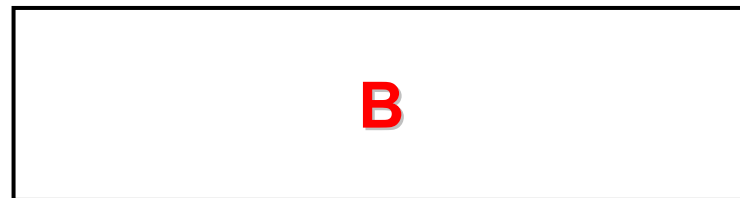
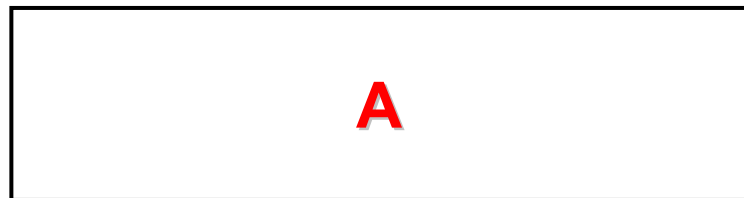
Диспетчер «глобальных ресурсов» LMD

- ✓ **«База данных» глобальных ресурсов** распределена между экземплярами Oracle9i RAC
- ✓ **Фоновый процесс LMD** каждого экземпляра управляет своей частью базы «глобальных ресурсов», таким образом, он является «центром управления» определенным «глобальным ресурсом»
- ✓ **Другие экземпляры** знают, LMD какого экземпляра является диспетчером для «глобального ресурса»
- ✓ **Если экземпляру** требуется глобальный ресурс в режиме, несовместимом с уже принадлежащим этому экземпляру, он отправляет запрос соответствующему LMD на преобразование глобальной блокировки
- ✓ **Получив уведомление** от LMD, экземпляр распоряжается «ресурсом» в соответствии с затребованным режимом блокировки
- ✓ **В случае сбоя экземпляра**, ответственность за «глобальные ресурсы» перераспределяется между LMD уцелевших экземпляров

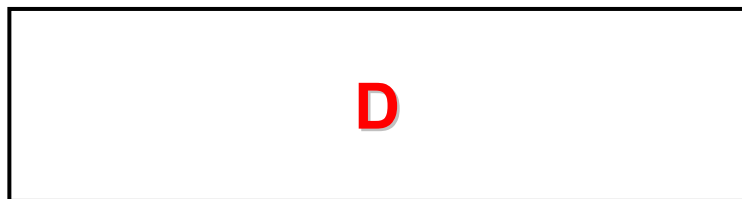
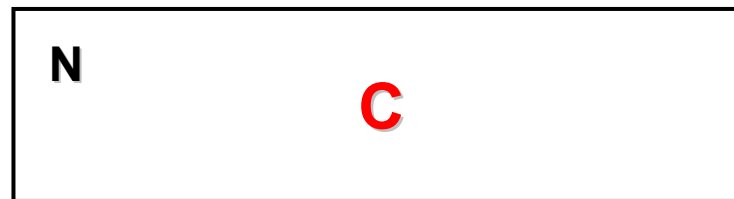
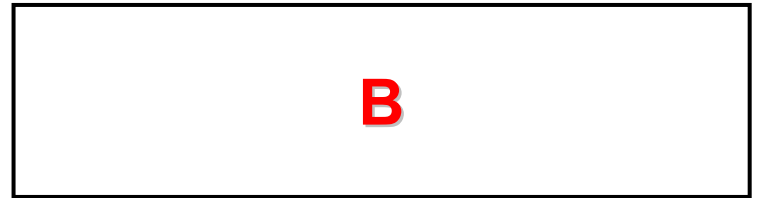
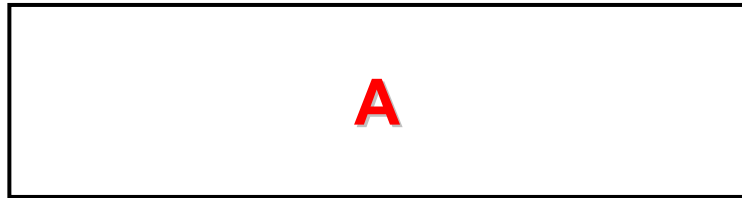
Cache Fusion

- ✓ **Механизм согласованной работы** с блоками базы данных в кеше буферов. Использует общий подход «глобальных ресурсов» и «глобальных блокировок»
- ✓ **«Глобальные ресурсы»** для Cache Fusion – это блоки БД
 - ✓ **Режимы**
 - ✓ **eXclusive** (монопольный)
 - ✓ **Shared** (разделяемый)
 - ✓ **Null** (никакой)
 - ✓ **Роли**
 - ✓ Локальная
 - ✓ Глобальная
 - ✓ **Состояние буферов (копий блоков БД в кеше буферов)**
 - ✓ **XCUR** (eXclusive CURrent)
 - ✓ **SCUR** (Shared CURrent)
 - ✓ **CR** (Current Read)
 - ✓ **PI** (Past Image)

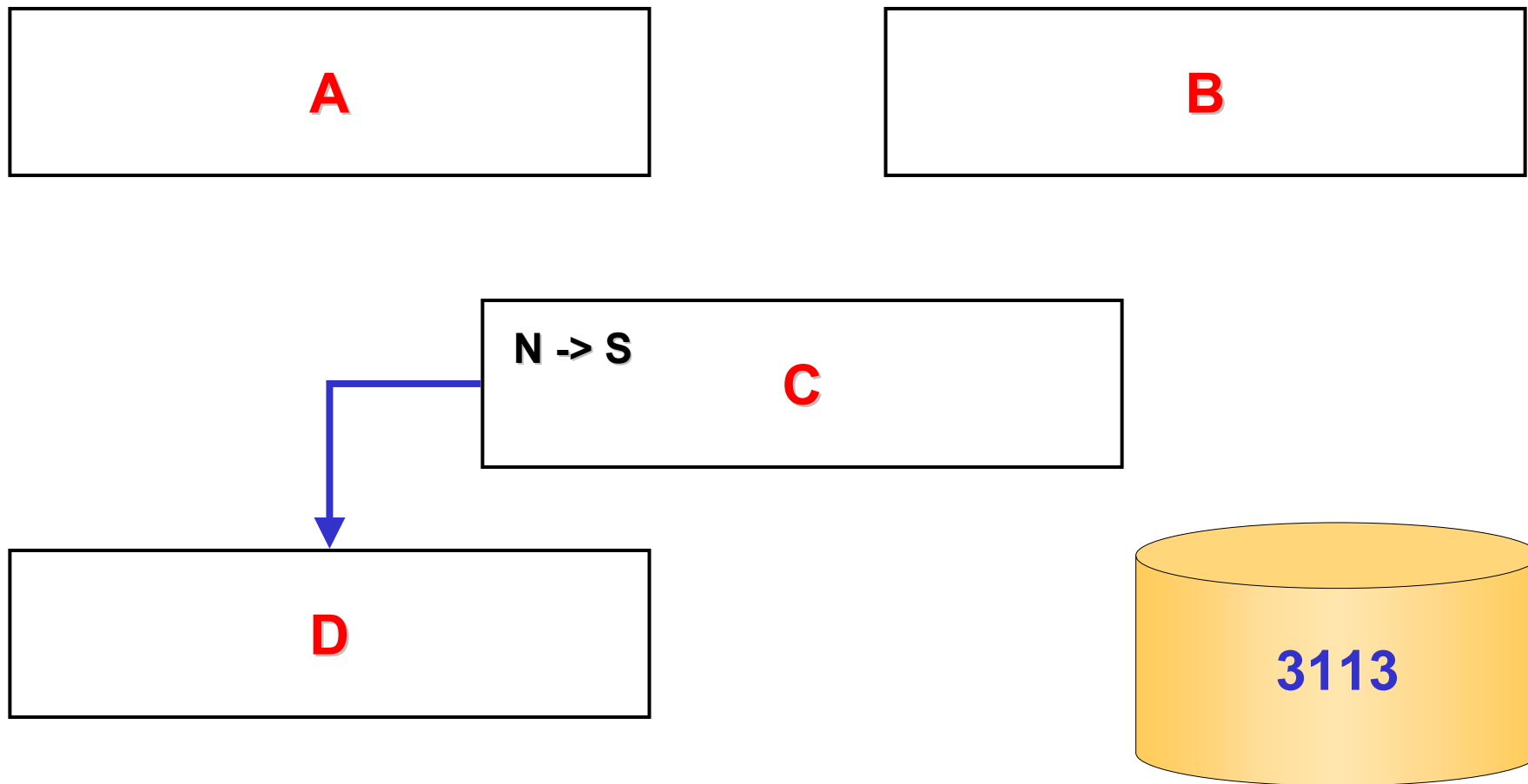
Cache Fusion – Схема примеров



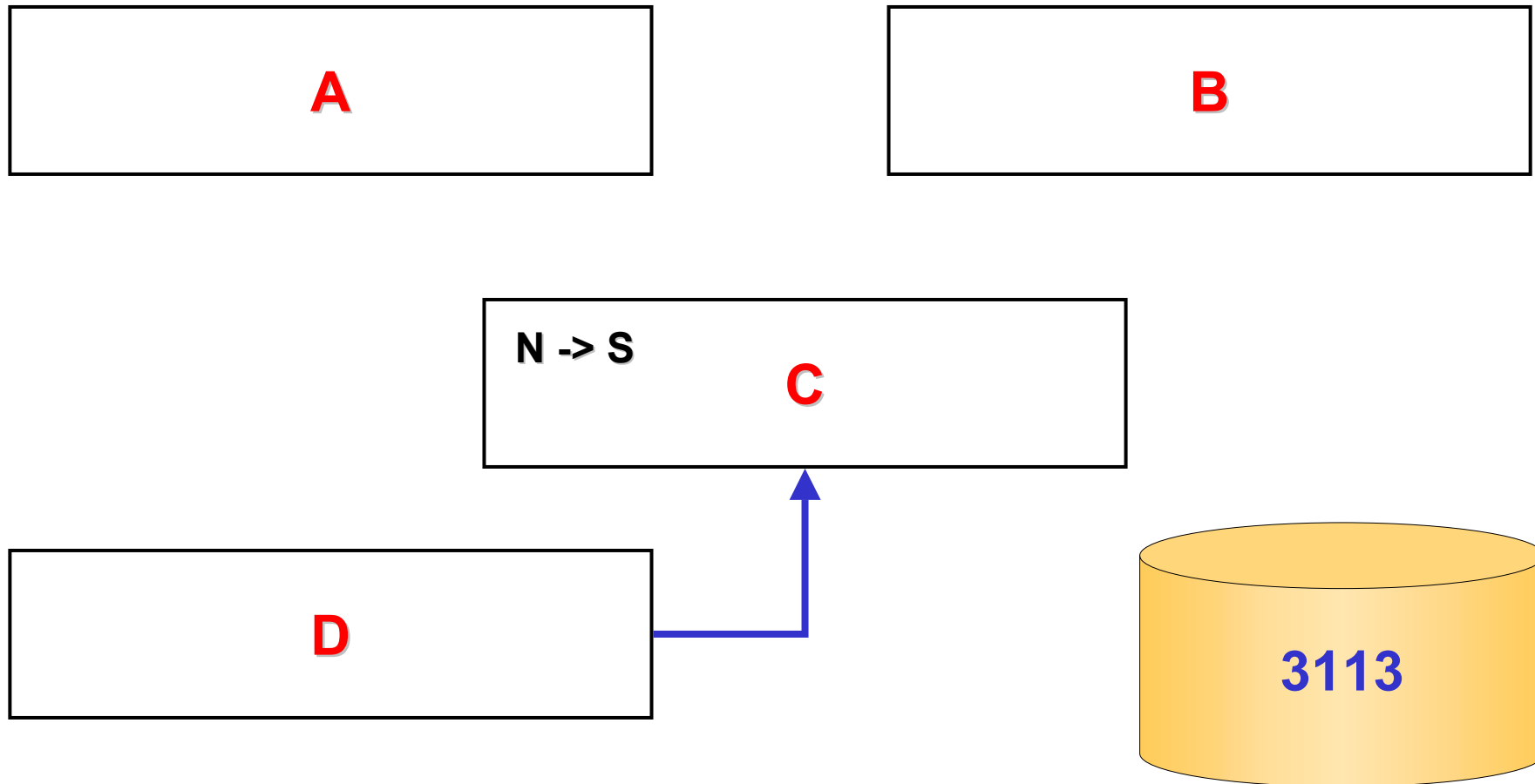
Cache Fusion: «Преобразование на чтение»



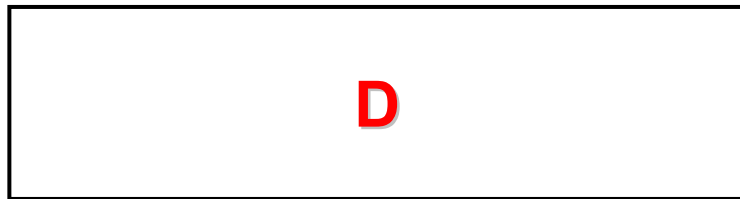
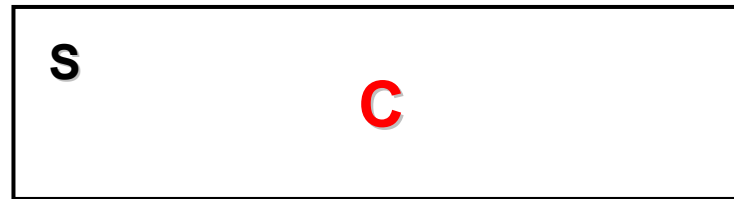
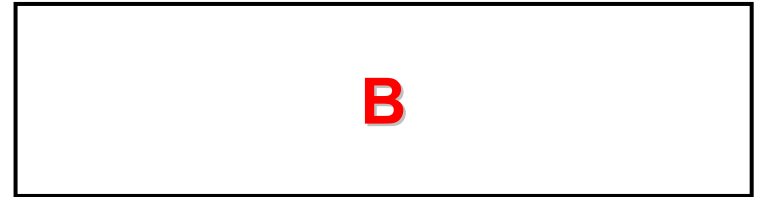
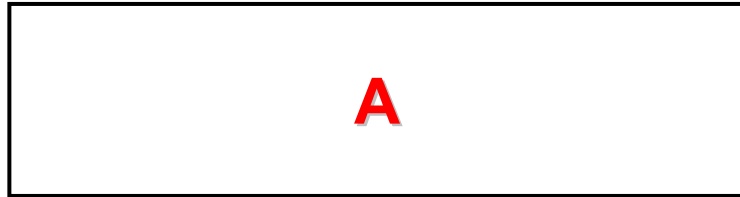
Cache Fusion: «Преобразование на чтение»



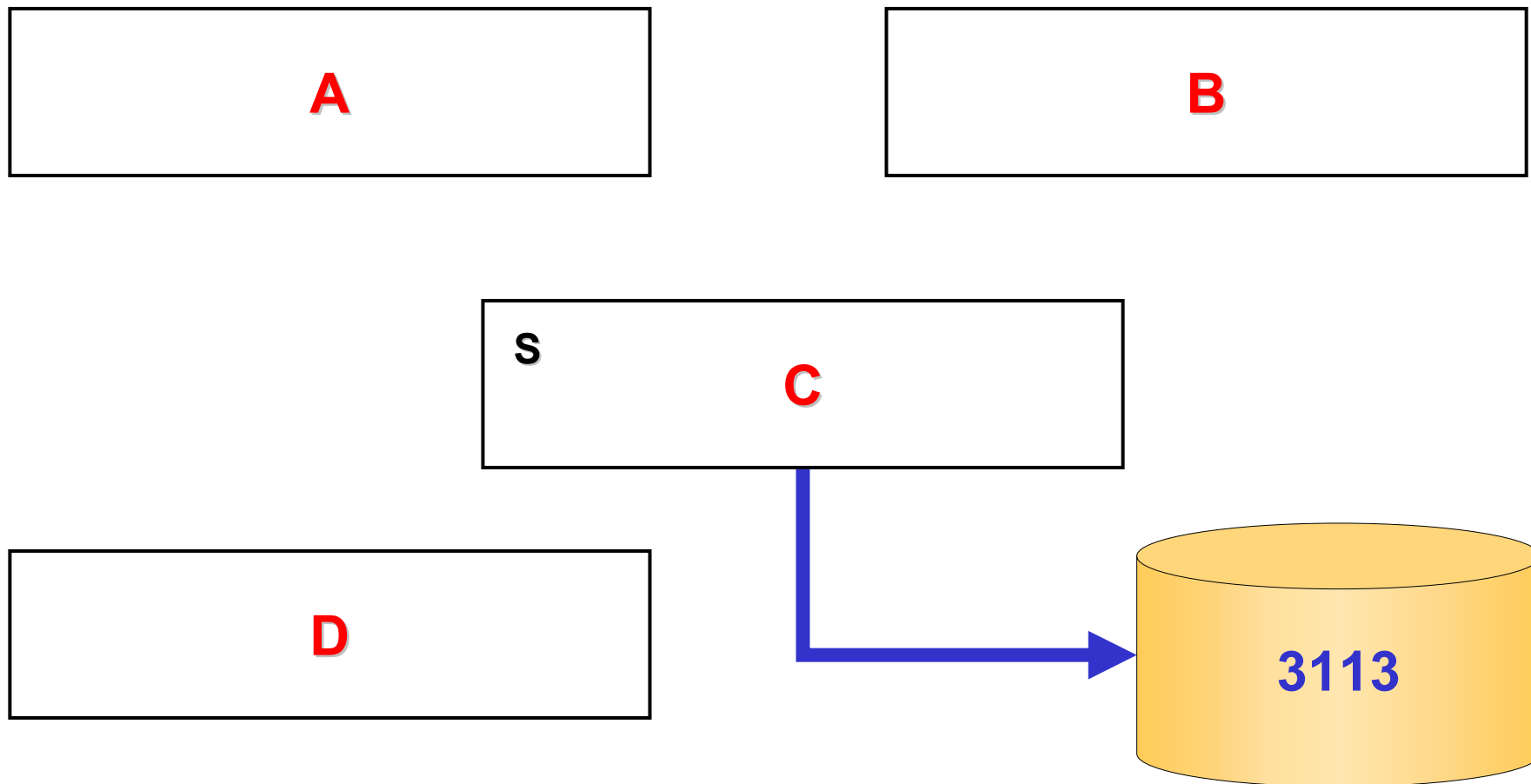
Cache Fusion: «Преобразование на чтение»



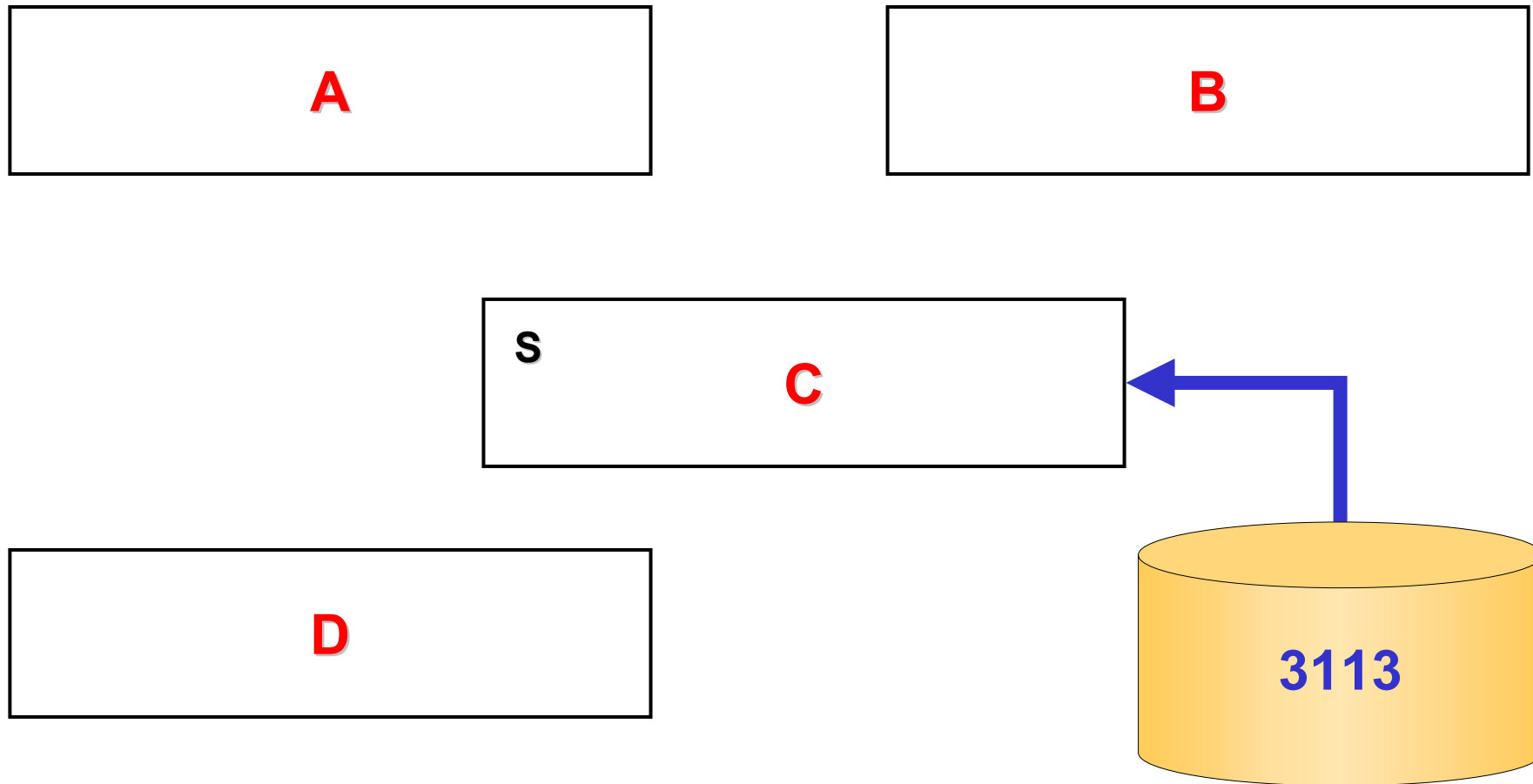
Cache Fusion: «Преобразование на чтение»



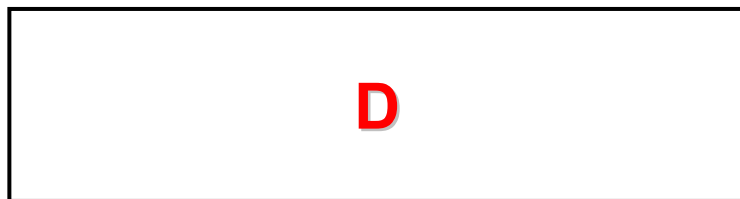
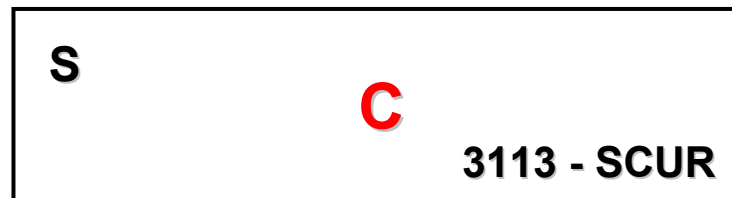
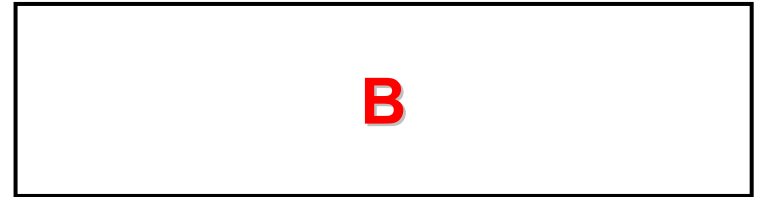
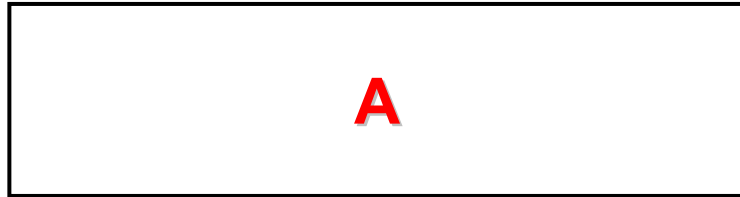
Cache Fusion: «Преобразование на чтение»



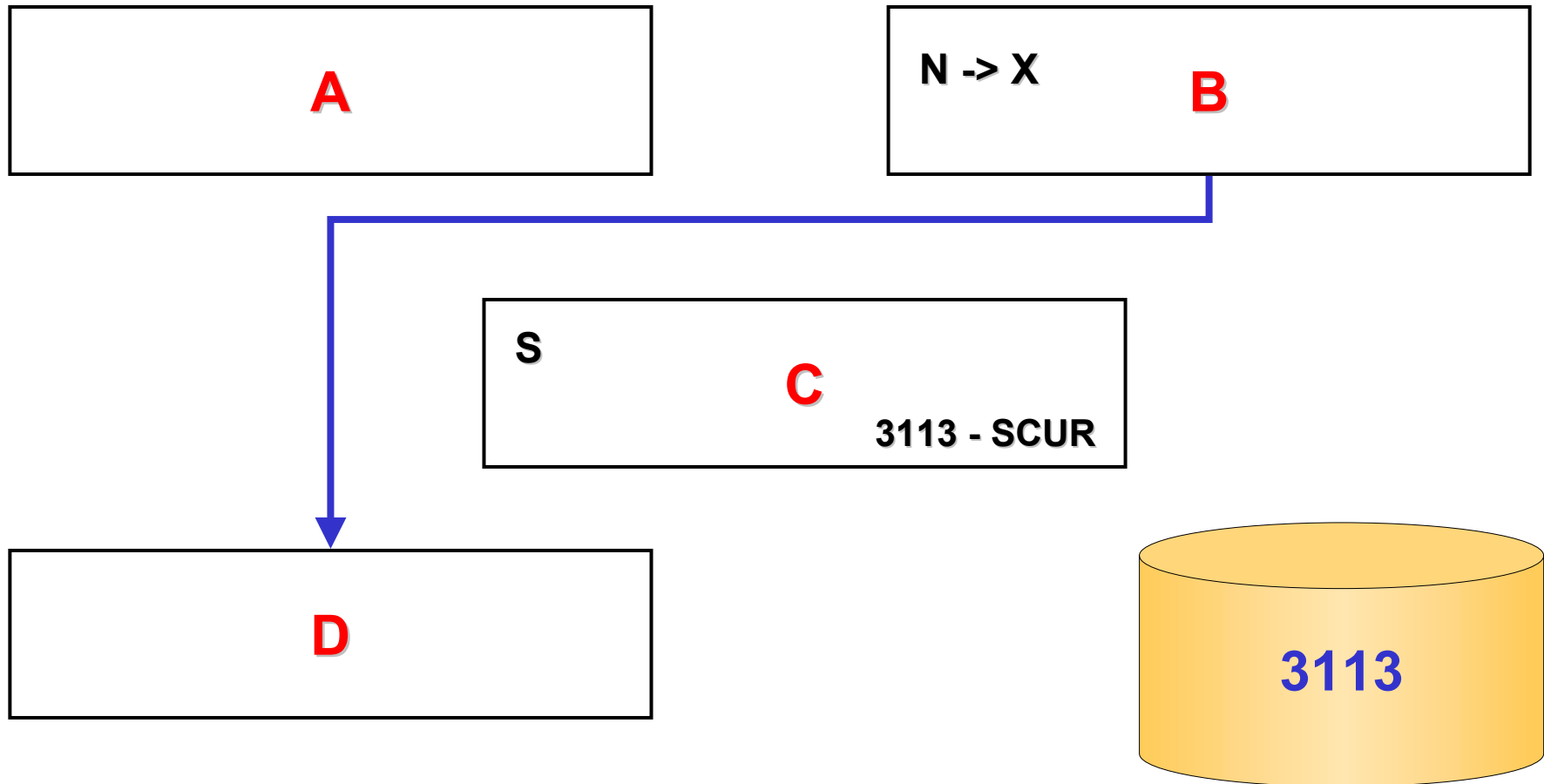
Cache Fusion: «Преобразование на чтение»



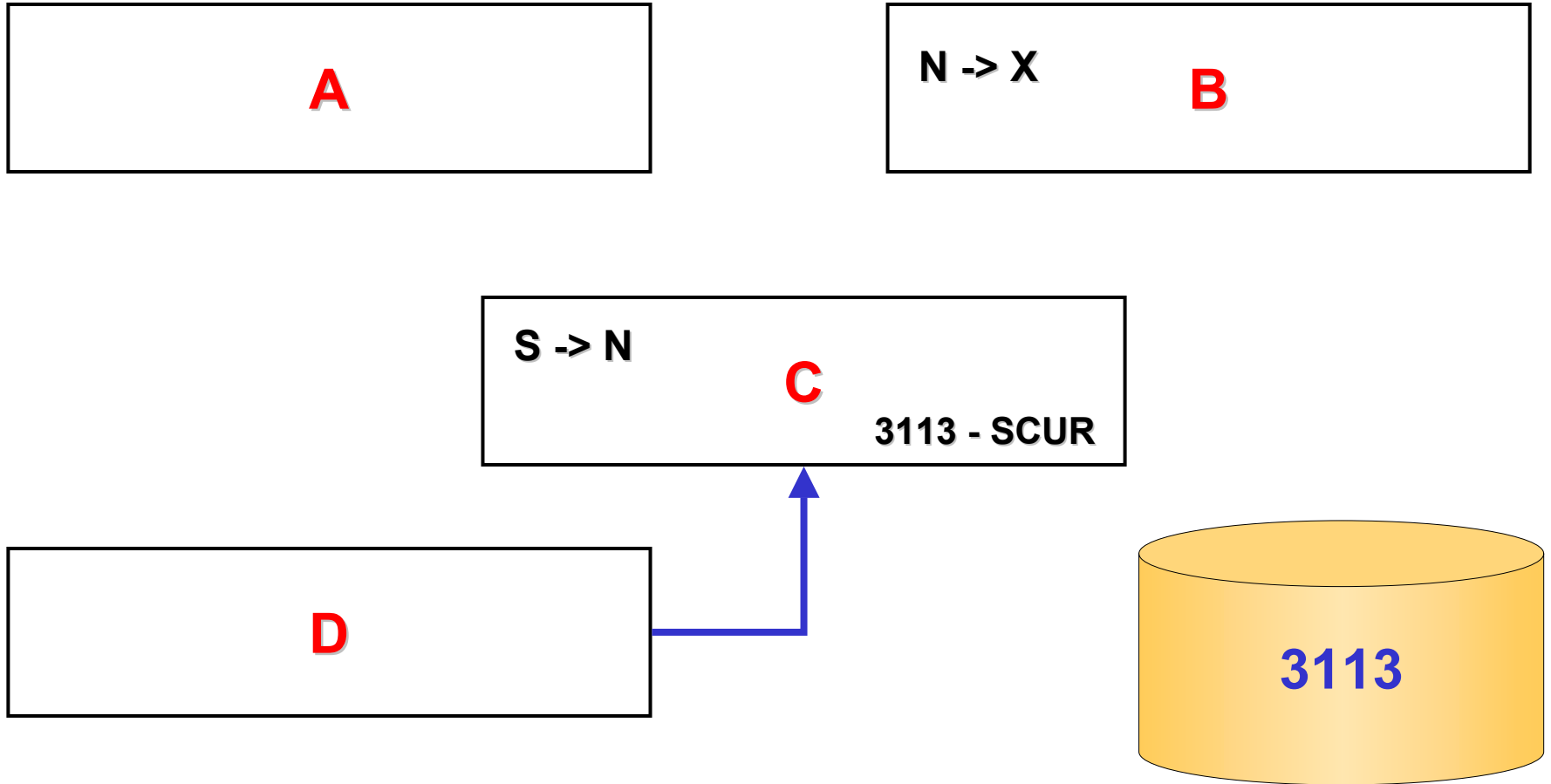
Cache Fusion: «Преобразование на чтение»



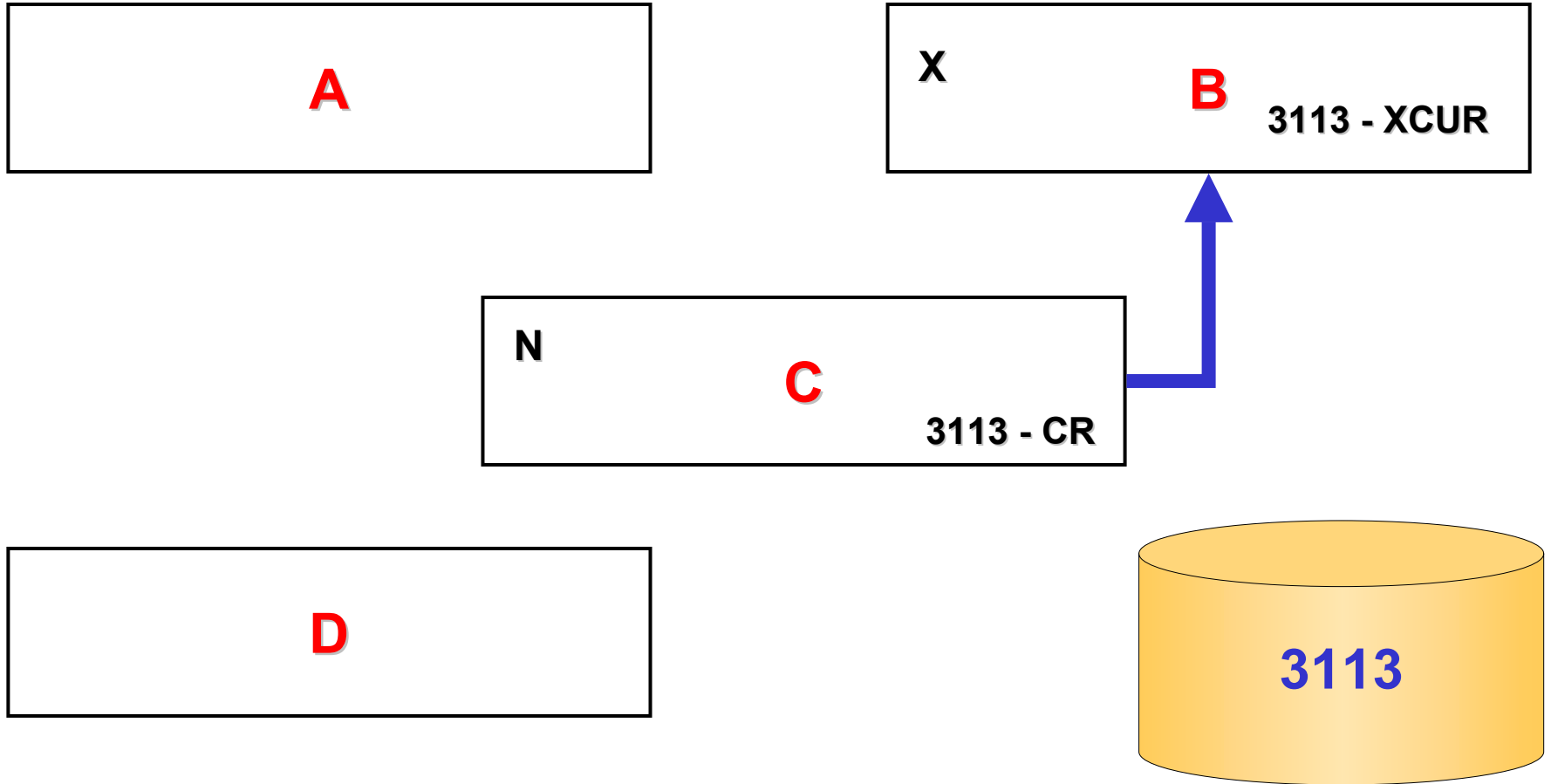
Cache Fusion: «Преобразование чтение - запись»



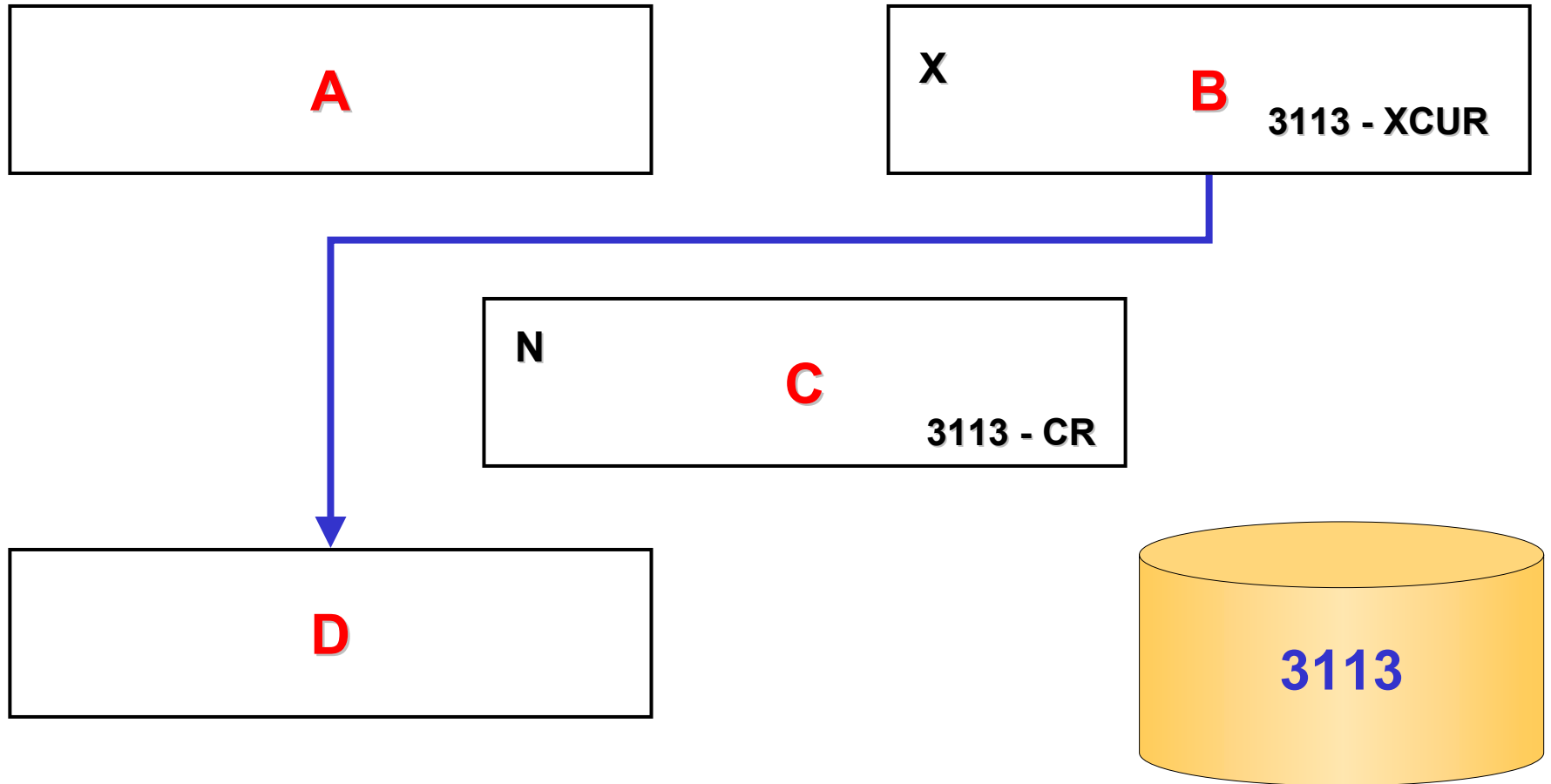
Cache Fusion: «Преобразование чтение - запись»



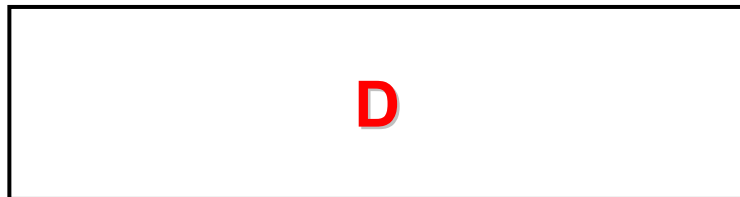
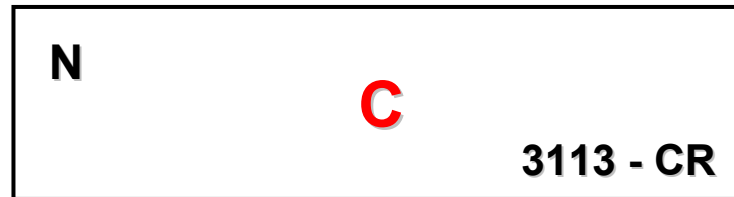
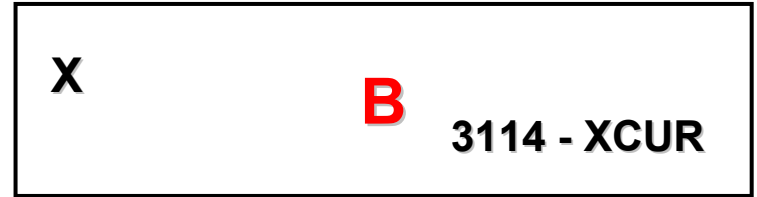
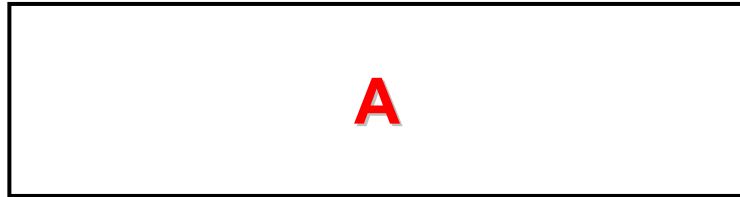
Cache Fusion: «Преобразование чтение - запись»



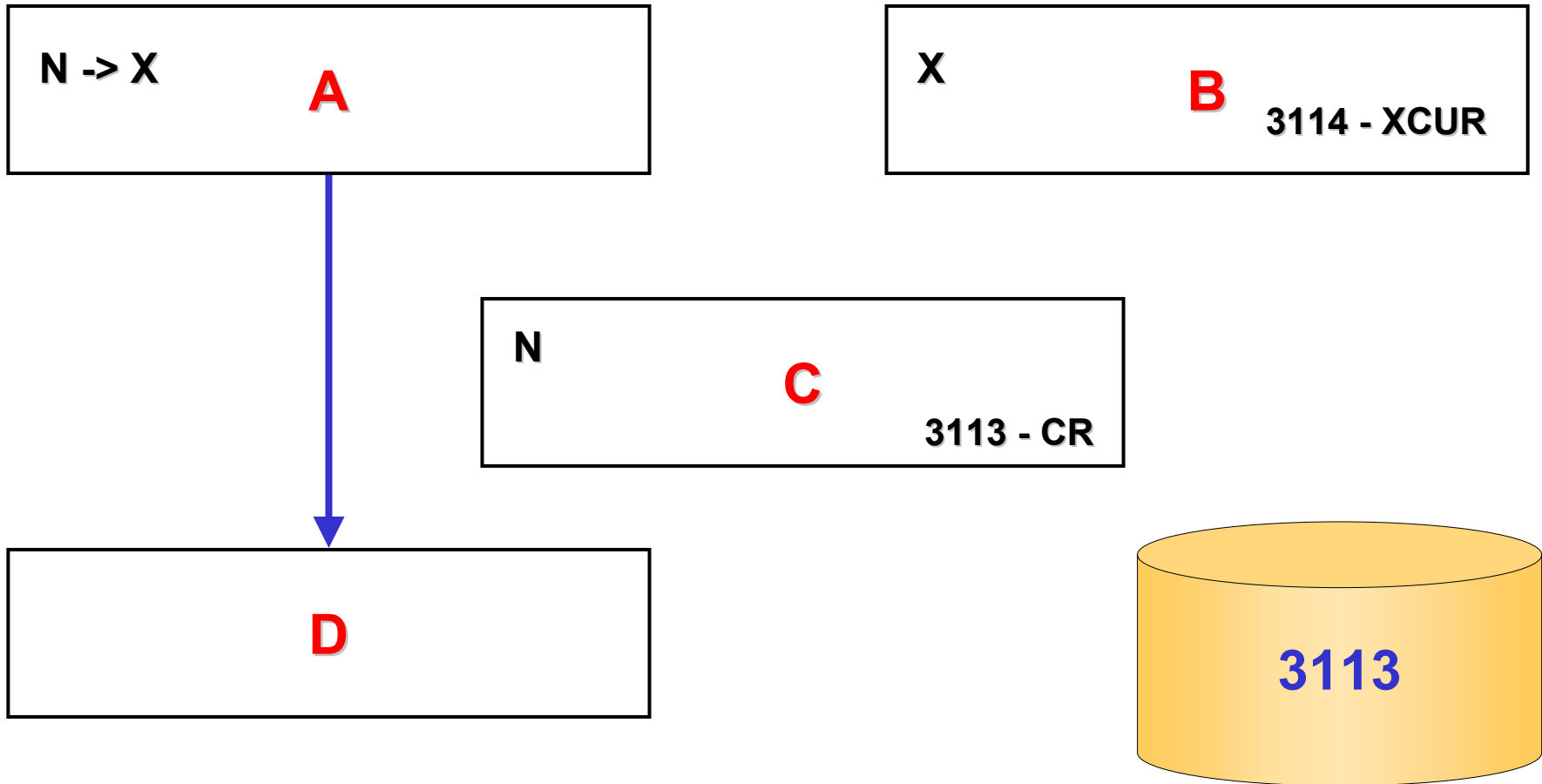
Cache Fusion: «Преобразование чтение - запись»



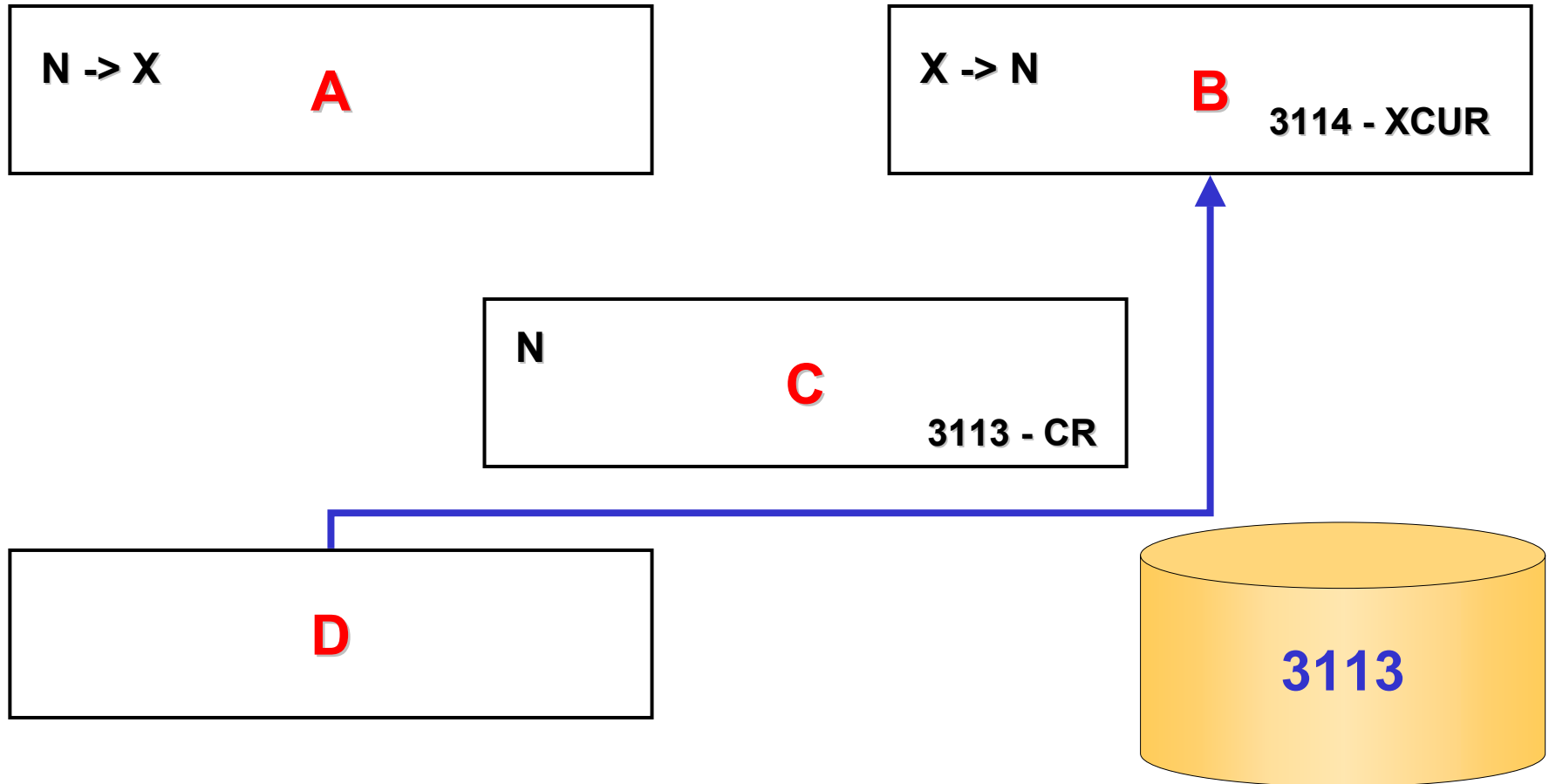
Cache Fusion: «Преобразование чтение - запись»



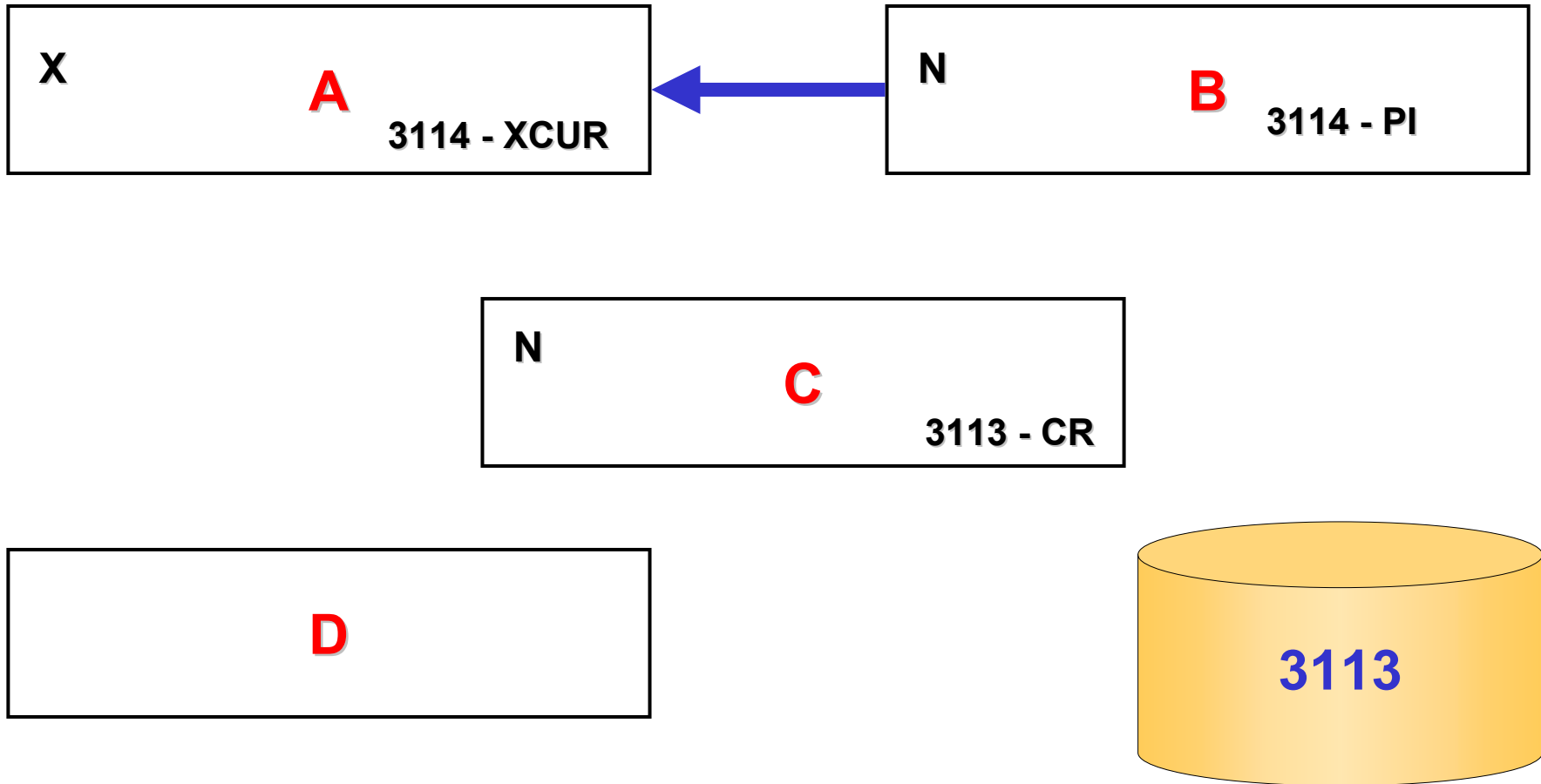
Cache Fusion: «Преобразование запись - запись»



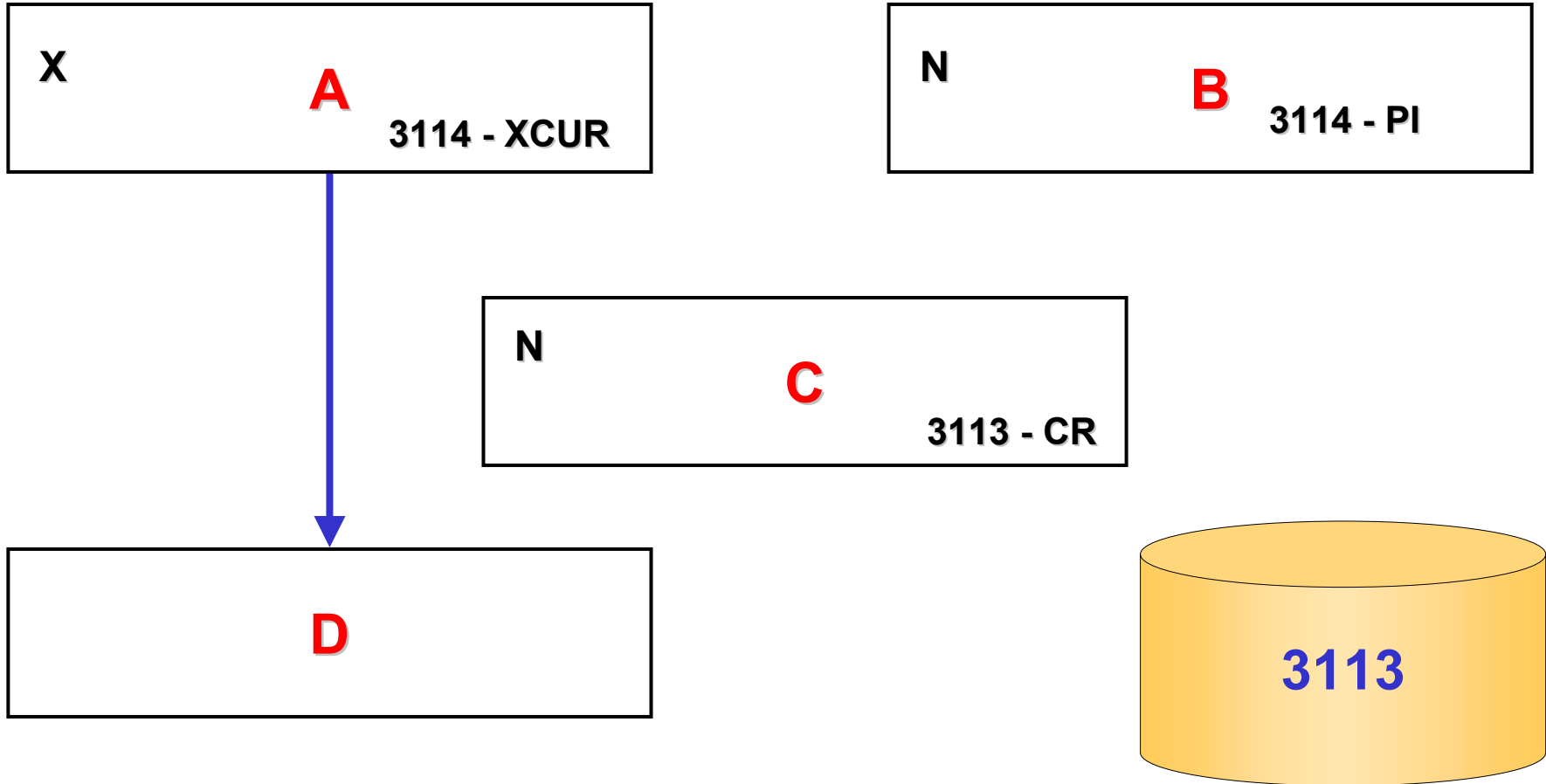
Cache Fusion: «Преобразование запись - запись»



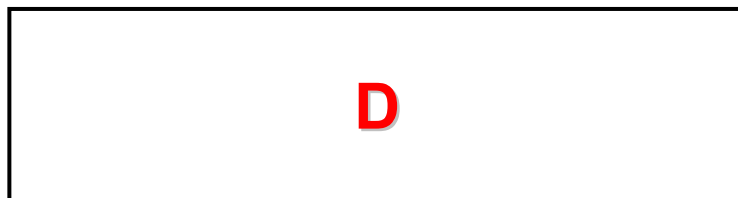
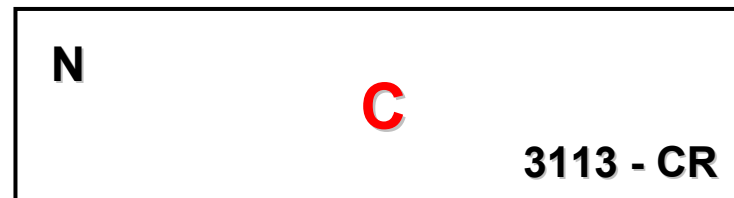
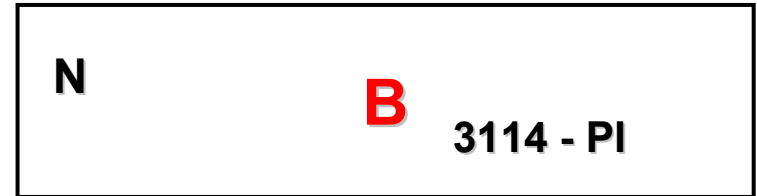
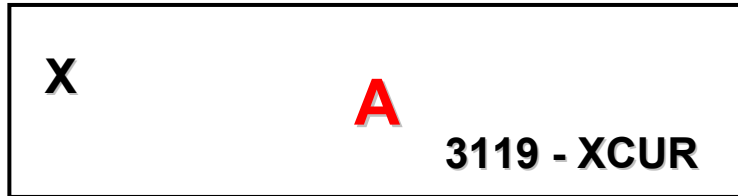
Cache Fusion: «Преобразование запись - запись»



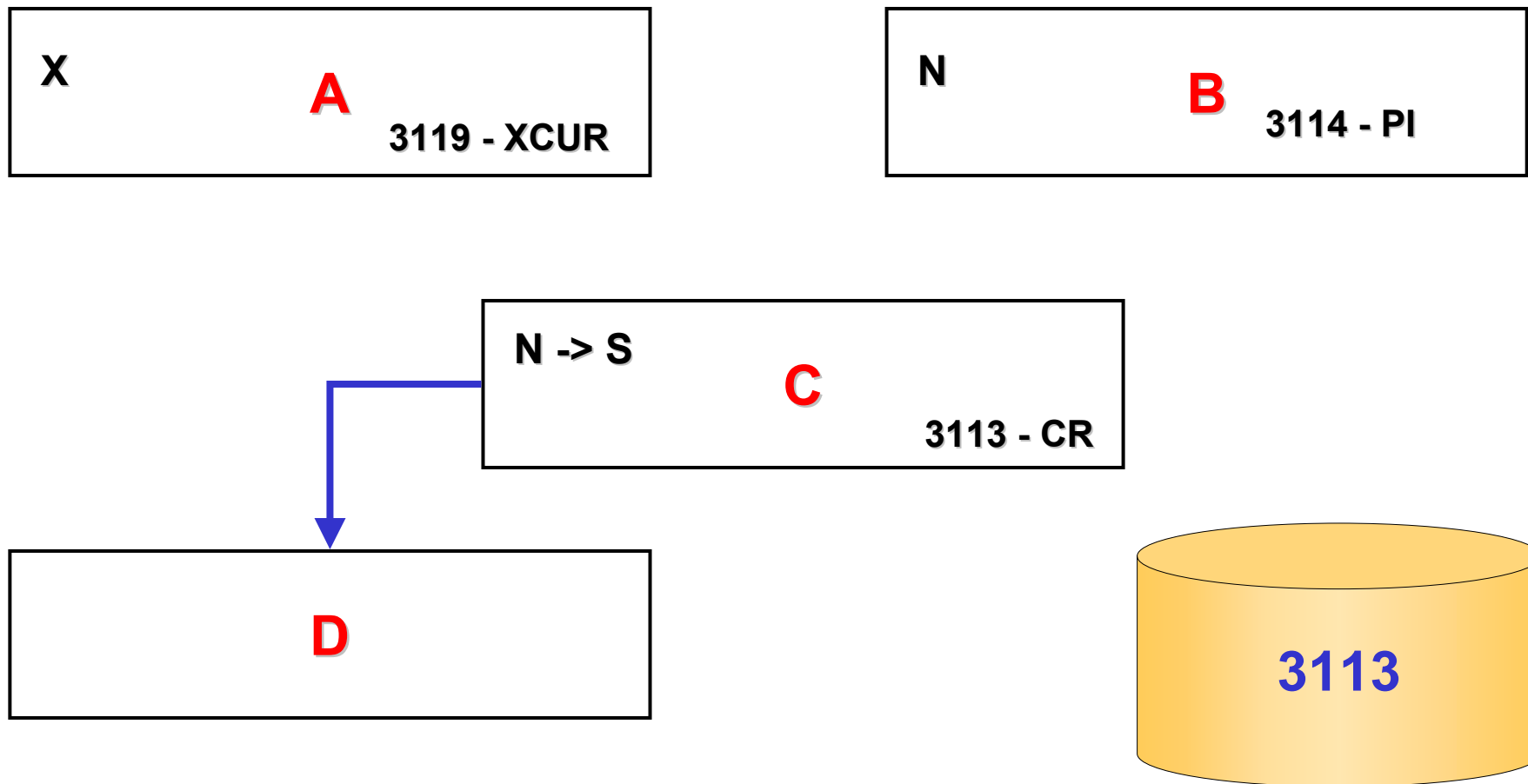
Cache Fusion: «Преобразование запись - запись»



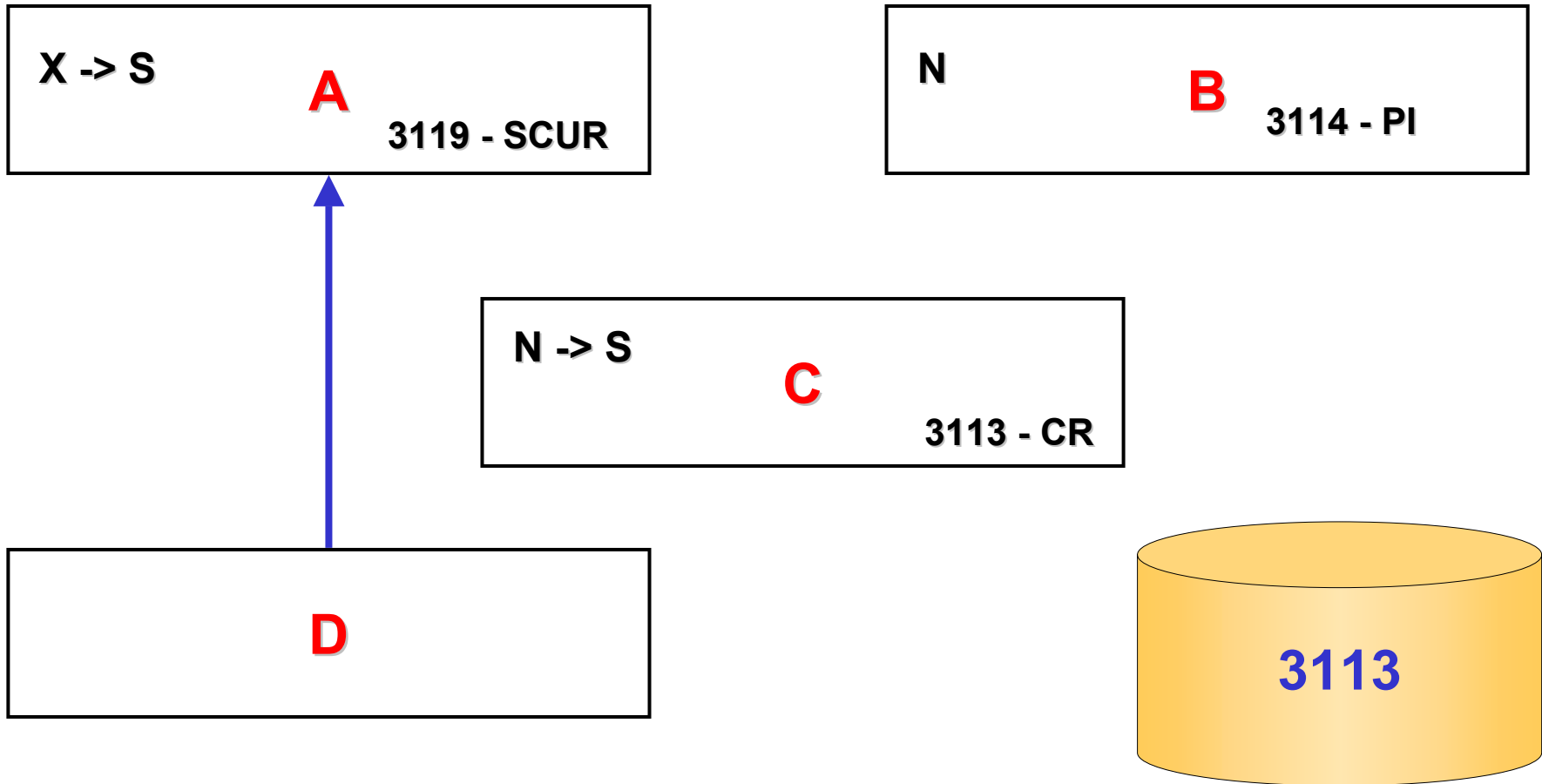
Cache Fusion: «Преобразование запись - запись»



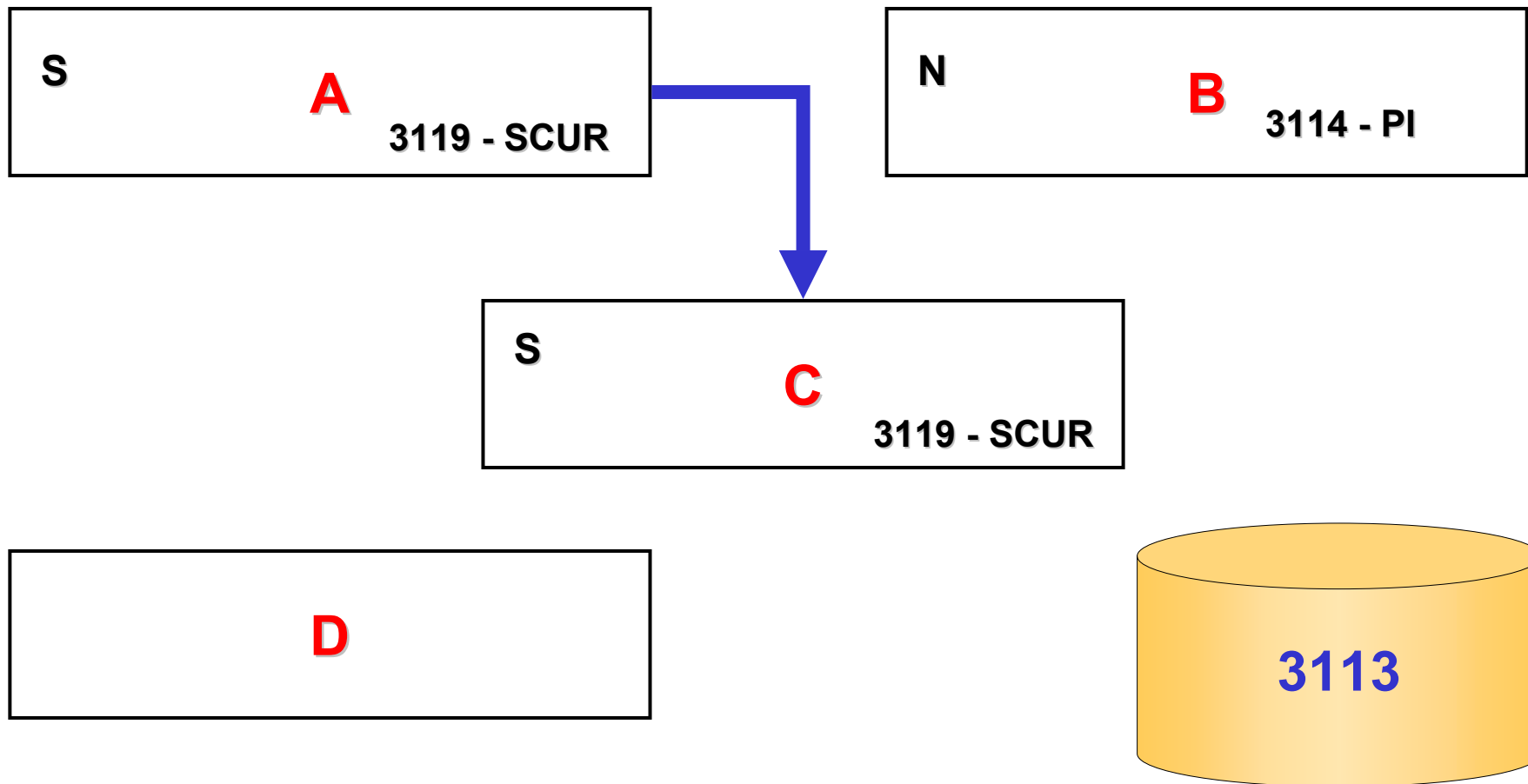
Cache Fusion: «Преобразование запись - чтение»



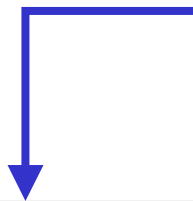
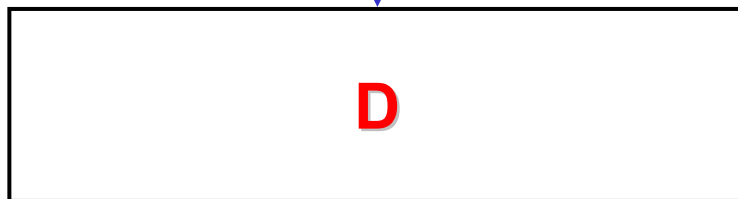
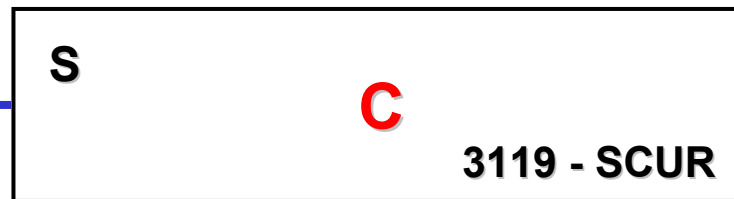
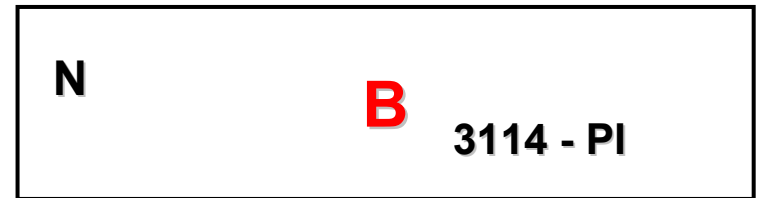
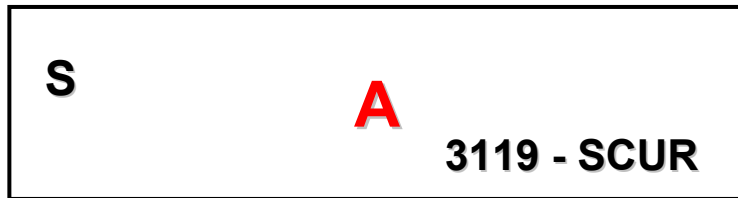
Cache Fusion: «Преобразование запись - чтение»



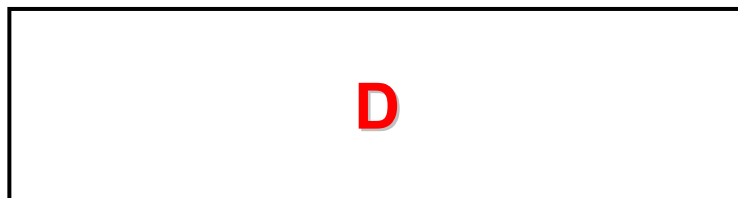
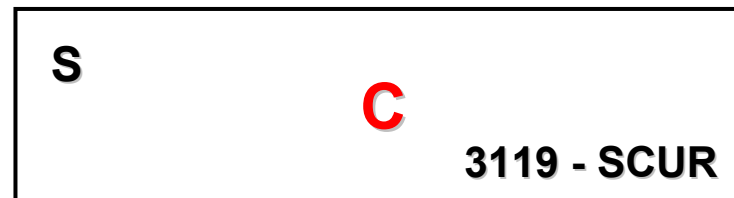
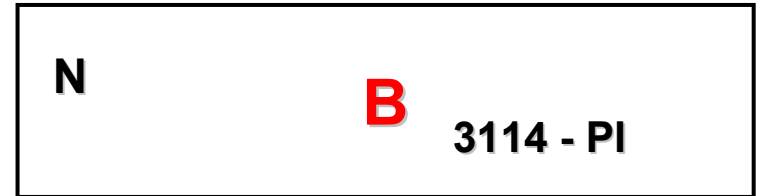
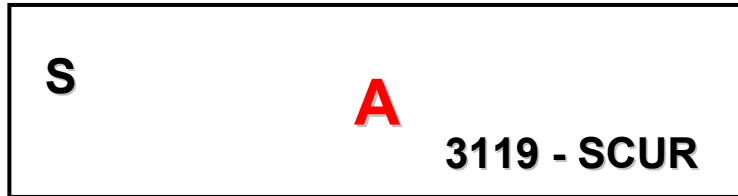
Cache Fusion: «Преобразование запись - чтение»



Cache Fusion: «Преобразование запись - чтение»



Cache Fusion: «Преобразование запись - чтение»



Глобальность блокировок и производительность

- ✓ **Для приложений типа «read only» (DSS/OLAP)** перенос в среду Oracle9i RAC практически не заметен
- ✓ **Для приложений типа «read/write» (OLTP)** возможны дополнительные накладные расходы из-за необходимости синхронизации кеша буферов
- ✓ **Влияние характеристик Interconnect'a существенно:**
 - Пропускная способность (bandwidth)
 - Latency (скорость доступа)
- ✓ **Дизайн приложения** может снизить требования к характеристикам Interconnect'a

Административные средства повышения производительности

- ✓ **Автоматическое управление пространством сегмента**
CREATE TABLESPACE ... EXTENT MANAGEMENT
LOCAL SEGMENT SPACE MANAGEMENT AUTOMATIC;
- ✓ **Альтернатива – использование FREELIST GROUPS**
- ✓ **Реверсивные индексы для сопровождения условий целостности**
CREATE INDEX ... ON ... REVERSE;
- ✓ **Кеширование последовательностей (sequences)**

РЕЗЮМЕ

- ✓ Oracle RAC является магистральным направлением в развитии серверных технологий Oracle.
- ✓ СУБД Oracle10g полагается на технологию Real Application Clusters
- ✓ Механизм «Cache Fusion» существенно снижает порог для масштабирования приложений типа OLTP
- ✓ Характеристики Interconnect'а существенно сказываются на производительности систем на базе Oracle RAC

Контактная информация

Андрей Криушин

Консультант по ПО Oracle
компании РДТЕХ

ORACLE | CERTIFIED
PROFESSIONAL

Andrei.Kriushin@rdtex.ru

Тел: +7(0967) 74 45 81, 74 08 76

Код из Москвы и области - 27

www.rdtex.ru

