Bases de datos avanzadas

Universidad de Los Andes

Postgrado en Computación

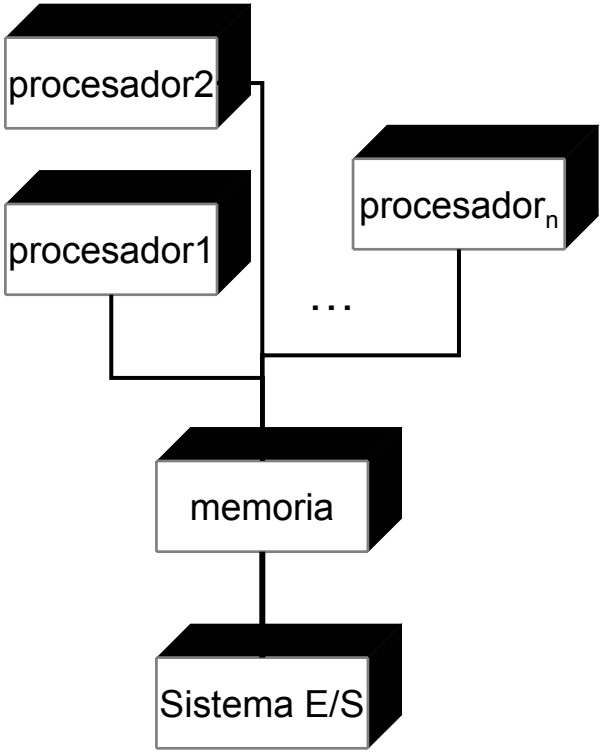
Prof. Isabel M. Besembel Carrera

Unidad II. Sesión 21. BD distribuidas y paralelas.

UNIVERSIDAD DE LOS ANDES

Bases de datos distribuidas

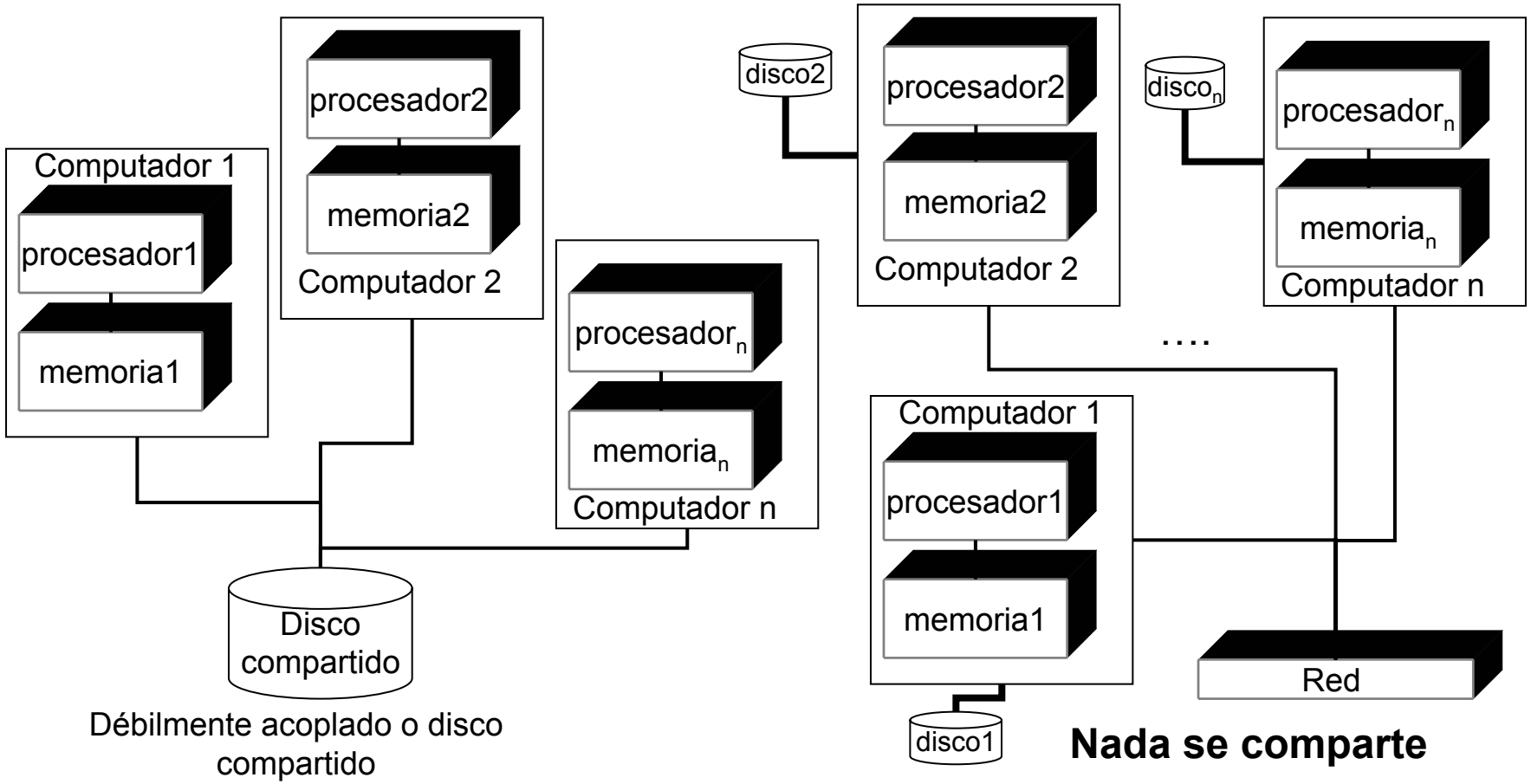
- Una BD distribuida es una colección de varias BD lógicamente inter-relacionadas sobre una red de computadores.
- Los usuarios tienen acceso integrado y transparente a una colección de BDs (1980).
- Actualmente una BDD es una colección de BDs independientes o federadas, donde cada sistema tiene facilidades para intercambiar datos y servicios con los otros miembros.
- Arquitectura: Fuerte y débilmente acopladas.



Compartir todo o fuertemente acoplado

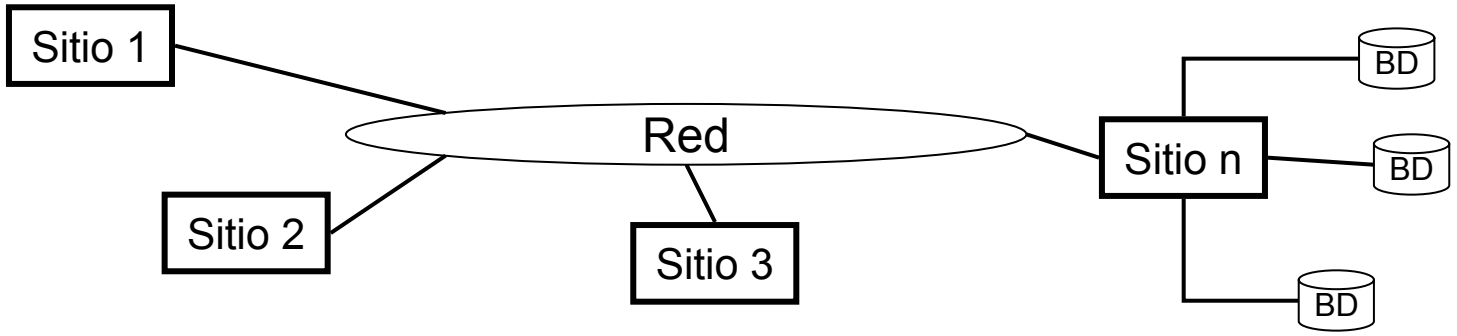
UNIVERSIDAD DE LOS ANDES

Arquitecturas



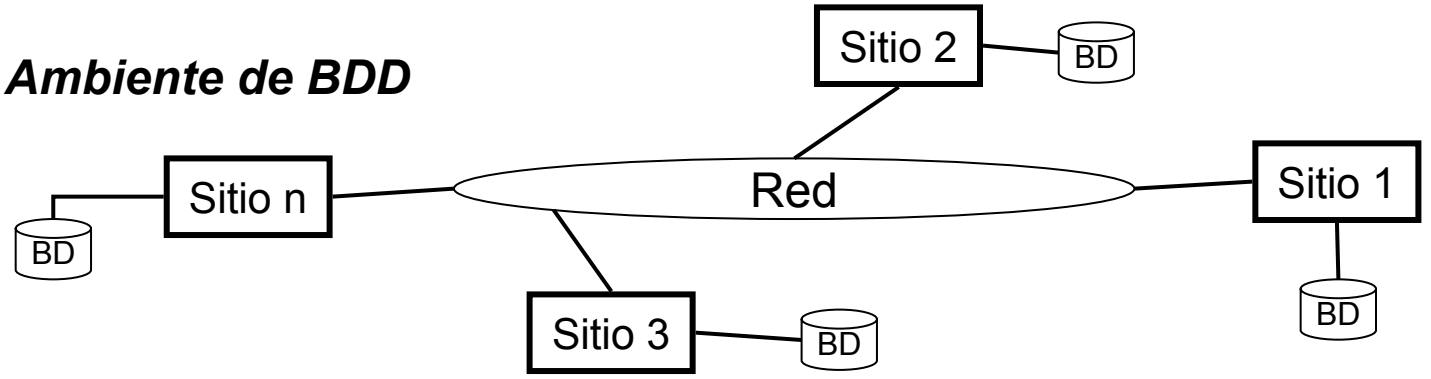
UNIVERSIDAD DE LOS ANDES

BD vs. BDD



BD central con acceso a red

Ambiente de BDD

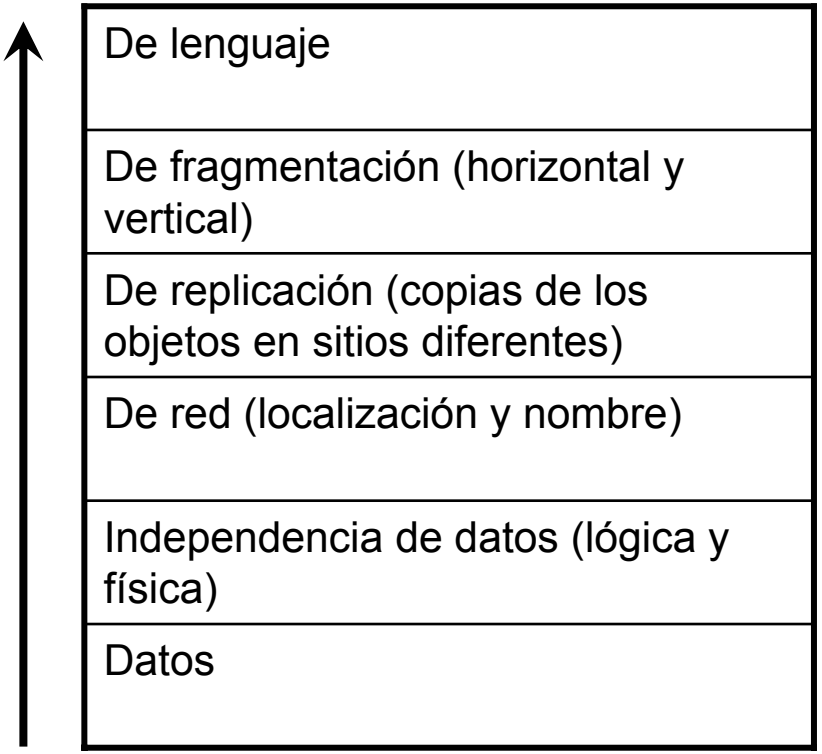


UNIVERSIDAD DE LOS ANDES

Ventajas y desventajas

- Ventajas:
  - ✓ Autonomía local
  - ✓ Mejora en el rendimiento
  - ✓ Mejora de la disponibilidad
  - ✓ Economía
  - ✓ Expansión
  - ✓ Se comparten datos
- Desventajas:
  - ✓ Falta de experiencia
  - ✓ Complejidad
  - ✓ Costo
  - ✓ Distribución del control
  - ✓ Seguridad
  - ✓ Cambios de centralizado a distribuido

Transparencias



UNIVERSIDAD DE LOS ANDES

Problemas

- Los sistemas de BDD pueden estar sobre diferentes plataformas de hardware, desde una colección de computadores débilmente acoplados conectados a una red de alta velocidad hasta un conjunto de máquinas geográficamente distribuidas conectadas por enlaces de bajo ancho de banda.
- Los datos pueden representarse en una variedad de modelos y pueden estar fragmentados y almacenados en computadores separados.
- *Problemas claves:*
  1. Para responder una consulta, los datos residentes en computadores separados deben ser combinados y movidos.
  2. Como los datos están bajo el control de computadores separados, es necesario coordinar las Ti para preservar la consistencia de datos.

UNIVERSIDAD DE LOS ANDES

Procesamiento de consultas

- El plan de ejecución dice la secuencia de operaciones y sus caminos de acceso
- Las BDD incluyen además, la construcción de un plan de ejecución que maximice el paralelismo y minimice la transferencia de datos en la red
  - ✓ Ejemplo: **Select * from** A, B **where** A.u > 100 **and** A.x = B.x
  1. Suponga que A está en N1 y B en N2. Debe hacerse una selección en N1 sobre A y luego un producto de A y B en N2.

     N1: **Select * from** A **where** A.u > 100

     poner el resultado en un *tubo* hacia N2

     N2: Recibir el resultado de N1 por el *tubo*

     realizar el producto de esas tuplas con B según x.
- En el mecanismo de *entubamiento* (pipelining) los nodos se conectan para la transferencia de tuplas y todos los nodos se utilizan concurrentemente en el procesamiento de consultas

UNIVERSIDAD DE LOS ANDES

Procesamiento de consultas

2. Suponga que A y B son muy grandes, pero que el resultado del producto A.x = B.x es relativamente pequeño.

   Paso 1: N2: Enviar (**Select** B.x **from** B) a N1 en la tabla T

   Paso 2: N1: **Select** A.* **from** A, T **where** A.x = T.x **and** A.u > 100

   entubar el resultado hacia N2

   N2: Recibir las tuplas de N1 por el tubo

   realizar el producto de esas tuplas con B según x.

- Este mecanismo de producto en dos pasos multilugar se denomina un semiproducto

UNIVERSIDAD DE LOS ANDES

Procesamiento de consultas

- Un sistema optimizador de consultas distribuido debe conocer el tamaño de las relaciones, el tamaño de sus atributos, los costos de transferencia y la distribución del esquema para maximizar concurrencia y minimizar la transferencia de datos.
- Ejemplo:
  - ✓ Suponga que A está fragmentada horizontalmente por x en m segmentos que están almacenados en NA1, NA2, ..., NAm;
  - ✓ y que B también está fragmentada de manera similar en NB1, NB2, ..., NBm.
  - ✓ Para acoplar A y B, las tuplas de A según x en NA1 solo necesitan ser comparadas con las de B en NB1, y así para las otras.
  - ✓ Así, NA1 debe ser entubado con NB1, NA2 con NB2, etc; y cada nodo NB debe enviar el resultado del producto al lugar de la consulta
- Hay otros factores que deben tomarse en cuenta para la optimización de las consultas distribuidas, como son: la carga de trabajo de los nodos, la capa física de conexión y los caminos de acceso y las propiedades deseadas para los resultados

UNIVERSIDAD DE LOS ANDES

Transacciones distribuidas

- En ambientes centralizados, las Ti se ejecutan con las propiedades ACAD: Atomicidad, Consistencia, Aislamiento y Durabilidad.
- Esto se asegura con el protocolo de validación (commit) en dos fases y el mecanismo de uso del diario
- En BDD la porción de una Ti en un lugar o computador se denomina una subTi y cada lugar debe asegurar que sus subTi sean ejecutadas atómicamente.
- Adicionalmente es necesario coordinar las decisiones hechas en cada lugar, para pasar todas las subTi o abortarlas todas
- El protocolo de validación dos fases realiza esta coordinación

UNIVERSIDAD DE LOS ANDES

Transacciones distribuidas

- Protocolo de validación en dos fases (2PC, Two Phases Commit)
  - ✓ Un lugar coordinador pide a todos los participantes la ejecución de las subTi.
  - ✓ Cuando cada participante termina, genera un estado de *preparado-para-pasar* donde es posible pasar la subTi o abortarla.
  - ✓ Envía un mensaje de preparado-para-pasar al coordinador, que al recibirlo desde **todos** los participantes, coloca una entrada de pasar en su diario y envía el mensaje de pasar a todos los participantes.
  - ✓ Cuando un participante lo recibe, él completa la subTi.
  - ✓ Si hay alguno que no puede o no quiere completar las subTi, lo comunica al coordinador, quien escribe una entrada de abortar en su diario y envía el mensaje de abortar a todos los participantes.
- El protocolo de bloqueo en dos fases requiere la realización de los bloqueos (compartidos o exclusivos) para todos los objetos que accede.
- Cada bloqueo de un objeto lo maneja el lugar donde está el objeto.

UNIVERSIDAD DE LOS ANDES

Transacciones distribuidas

- Cuando hay replicación, todas las copias de un objeto X deben ser bloqueadas adecuadamente.
- Desventajas: Para bloquear X deben bloquearse todas sus copias y si alguna de ellas no está disponible, la Ti no puede ser hecha
- Para reducir la redundancia de bloqueos y los problemas de disponibilidad, se tiene como meta la de garantizar una copia semántica, esto es como si existiera una sola copia.
- Algunos protocolos permiten solo bloquear la copia primaria, otros bloquear solo las copias disponibles, otros que no todos los participantes con copia pasen, etc.
- Actualmente no todos los manejadores proveen las facilidades de optimización de consultas distribuidas o de manejo de réplicas

UNIVERSIDAD DE LOS ANDES

Capacidades de los SGBD comerciales

- Muchos ofrecen distribución basada en la arquitectura cliente-servidor, donde muchas aplicaciones cliente en nodos cliente están conectadas a un servidor de BD en un nodo servidor. Ejm: Ingres/Net (1983).
- Además de los servidores de BD por SQL, otros ofrecen procedimientos almacenados y ejecutados en el servidor.
- Estos procedimientos generalmente efectúan una secuencia de consultas y regresan sus respuestas a través de un procedimiento remoto. Ellos son: Sybase, Informix, Oracle, DB2, etc.
- Ellos logran aislar las BD clientes de las estructuras y lenguajes soportados en la BD servidor, aumentando la autonomía y formando una base para la integración de BD servidoras heterogéneas.
- Protocolos estándares: SQL Access Group's, SQL Access API y Message Format

UNIVERSIDAD DE LOS ANDES

Capacidades de los SGBD comerciales

- Un cliente puede estar conectado simultáneamente a varios servidores de BD.
- La correlación de los datos de varias fuentes es responsabilidad del cliente.
- Un SMBDD ofrece las facilidades para manejar la transparencia de red y manejo distribuido de Ti.
- Un cliente envía sus peticiones al Monitor de Procesamiento de Ti (MPT) que usa el protocolo 2PC (Ingres de 1990, Informix versión 5.0 y Oracle release 7) Sybase no ofrece C2F automático.
- La mayoría lo ofrece sólo para ambientes homogéneos, todos los SMBDD del mismo fabricante
- El protocolo estándar XA y el ISO/TP permiten varios manejadores de Ti (Informix usa el XA desde 1991 y Oracle release 7 también)
- El enfoque teórico usado por los investigadores para soportar la transparencia de red fue el de arriba-hacia-abajo, mientras que los productos comerciales usan el de abajo-hacia-arriba

UNIVERSIDAD DE LOS ANDES

Capacidades de los SGBD comerciales

- Una BDD residente en varios lugares se comunica con un servidor de integración que ofrece un esquema global para las múltiples BDDs locales.
- Ingres/Star ofrece la transparencia de red a través de su servidor de integración, que distribuye y descompone la consulta sobre el esquema global en subconsultas sobre los esquemas locales soportados por el servidor de BD.
- Las BD locales permanecen sin conocer las otras BD locales y se comunican solamente a través del servidor de integración.
- Una BD local puede participar en varios servidores de integración y así participa en varios esquemas globales
- El procesamiento de consultas distribuido utiliza los parámetros de distribución estadística de los datos, cantidad transferida de datos, velocidad de los canales de comunicación, la existencia de índices primarios y secundarios y algunos parámetros de las estructuras de almacenamiento. Ejm: Ingres/Star

UNIVERSIDAD DE LOS ANDES

Capacidades de los SGBD comerciales

- El soporte de replicación aún es primitivo.
- La forma como se soporta es por medio de una fotografía de la BD, que genera y refresca una copia de solo lectura de los datos especificados en un intervalo de tiempo definido.
- Una fotografía de la BD es una vista materializada de la misma.
- Ejemplo:

  Create materialized_view v1 **as**

  select * from t **where** at = me

  refresh incremental cycle 1 hour

- Oracle y Rdb ofrecen esta facilidad

UNIVERSIDAD DE LOS ANDES

Investigación

- La transparencia puede ser ofrecida a varios niveles.
  - ✓ Transparencia completa: los lugares deben ponerse de acuerdo en el modelo de datos, la interpretación del esquema, la representación de datos, las funcionalidades disponibles y donde se localizan los datos.
  - ✓ Si no hay tal transparencia: solo debe haber acuerdo en el formato de intercambio de datos y las funciones proveídas por cada lugar
- Transparencia completa es lo que desea un usuario y lo que hace más difícil el control desde el punto de vista del administrador, dados los mecanismos de seguridad actuales.
- El modelo de servicio (sin completa transparencia) encapsula los datos permitiendo el acceso sólo a través de los procedimientos, lo que asegura la consistencia, necesita menos control de los participantes y no necesita un operador de producto multilugar, permitiendo mayor autonomía.
- Un modelo intermedio entre los anteriores no ha sido bien definido

UNIVERSIDAD DE LOS ANDES

Investigación

- El modelo de Ti convencional, 2PC, forza a los participantes a usar el mismo modelo y lo hace perder su autonomía, ya que uno es coordinador y los otros subordinados.
- El problema principal es de escala.
  - ✓ Cuando el número de participantes crece o la cantidad de trabajo crece, el tiempo de uso de los recursos también crece
- Un modelo de Ti no global como las *sagas* permite dar más autonomía y mejorar el rendimiento garantizando algunas propiedades de correctitud básicas.
- Una saga es una secuencia de pasos en varios lugares.
  - ✓ Solo se aseguran las restricciones de consistencia locales.
- Ejemplo: Si X en A es una copia de Y en B, es posible mantener una copia aproximada

UNIVERSIDAD DE LOS ANDES

Investigación

- El crecimiento de las BD ha puesto en relieve que los algoritmos de los SMBDD no se adecuan bien al número de componentes cuando el sistema crece, el tiempo de reorganización o de respaldo crece, la noche no es suficiente.
- Si la BD crece en alcance a nivel mundial, entonces allí no hay noches.
- Tener un directorio central de todos los recursos es impráctico por su gran tamaño, porque es más propenso a fallas y porque no todos los lugares pueden querer publicar sus recursos a los otros.
- Por ello el problema de encontrar los recursos en una BDD muy grande es un problema no resuelto. Ejemplo: búsquedas bibliográficas.
- Otros problemas: la detección de interbloqueos (abrazo mortal), el procesamiento de consultas y de Ti en sistemas grandes o en sistemas con consultas complejas o con muchos participantes por Ti

UNIVERSIDAD DE LOS ANDES

Investigación

- Problema de la administración de BDD grandes:
  - ✓ el volumen de cuentas,
  - ✓ autenticación de usuarios,
  - ✓ evaluación y entonación de la BDD,
  - ✓ instalación y actualización de paquetes,
  - ✓ manejo de los enlaces,
  - ✓ protocolos y rendimiento de la red, etc.
- Computación móvil: advenimiento de los laptops y comunicaciones por celular, complica aún más el manejo de BDD
  - ✓ Una colección de computadores móviles conectados por enlaces de radio. Ejm: Taxis con computadores y radio. ¿Cuál es el taxi más cercano?

UNIVERSIDAD DE LOS ANDES

Investigación

- Problemas:
  - La potencia eléctrica del computador es limitada, por lo que hay que minimizar el número de bytes transferidos, las consultas y las actualizaciones.
  - Además el computador del lugar es apagado frecuentemente para ahorrar energía, lo que hace que se necesiten mecanismos de reinicio eficientes y rápidos, así como también mecanismos de actualización de datos o copias por desconexión

- Aplicaciones no tradicionales:
  - BD comerciales para ventas por computador
  - Redes de información como MiniTel
  - Periódicos o revistas electrónicas
- La alternativa es proveer al usuario con una caja de herramientas para el acceso de la información adecuada a cada BDD que necesite usando una única interfaz y un único protocolo de comunicación

UNIVERSIDAD DE LOS ANDES

BDR paralelas

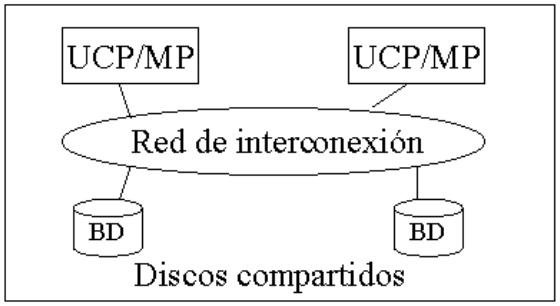
- Recursos:
  - ✓ procesadores, módulos de memoria principal y almacenamiento secundario (discos).
  - ✓ Dependiendo de la interacción de estos recursos, se definen varias arquitecturas de máquinas con varios procesadores
- Una BDRP se diferencia de una BD Distribuida que está bajo una red de área local en que la ***BDRP no tiene noción de la autonomía del lugar, tiene un esquema centralizado y tiene un único punto de inicio para la ejecución de todas sus consultas***
- Motivación:
  - ✓ El tamaño de las BDR varia desde los GB a TB
  - ✓ Rendimiento de los computadoras paralelos

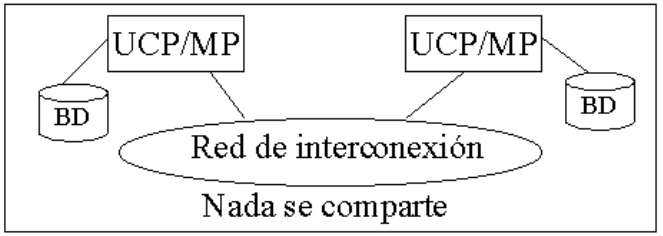
UNIVERSIDAD DE LOS ANDES

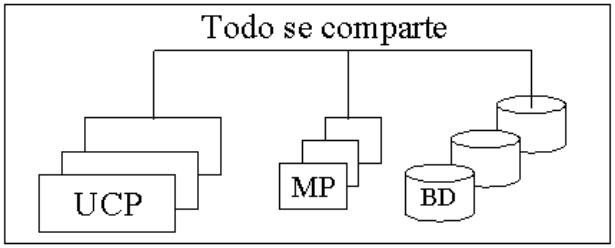
BDR paralelas

- ✓ Necesidad de altos rendimientos en las BDR muy grandes
- ✓ Los sistemas multiprocesadores de alto rendimiento y de propósito general ofrecen la facilidad de usarlos para procesar BD que permiten minimizar las E/S

- Arquitecturas:
  - ✓ **Discos compartidos:** Cada procesador tiene su propia MP, pero comparte su MS con todos
  - ✓ Todo se comparte: Todos los procesadores acceden una MP común y todas las MS (discos)
  - ✓ Nada se comparte: Cada procesador tiene su MP y MS, pero se comunican con pase de mensajes







UNIVERSIDAD DE LOS ANDES

SGBDRP

- DBC/1012 (Teradata Corp.-85)
  - ✓ *Consiste de un conjunto de procesadores de interfaces (PIF), de módulos de acceso (PMA) y las unidades de disco (UD).*
  - ✓ *Los PIF se comunican con la anfitriona, optimizan, difunden los datos, acumulan los resultados y dirigen la ejecución de los requerimientos de los usuarios.*
  - ✓ *Los PMA se encargan de almacenar y recuperar los datos de las UDs.*
  - ✓ *PIF y PMA se conectan por una estructura de árbol llamada Ynet, que es un bus activo capaz de realizar selecciones y ordenamientos.*
- Gamma (Universidad de Wisconsin-90)
  - ✓ *32 procesadores iPSC/2 Intel formando un hipercubo con 1 disco en cada nodo. Las relaciones se dividen horizontalmente sobre todos los discos del sistemas (desagrupamiento), según 4 métodos.*

UNIVERSIDAD DE LOS ANDES

SGBDRP

- Bubba (MCC-90)
  - *Basado en el modelo de nada se comparte e implementado en 40 nodos Flex/32 procesadores donde cada uno tiene su MS. Utiliza 3 tipos de nodos: procesadores de interfaz, depósitos inteligentes y depósitos inteligentes de puntos de chequeo y bitácoras.*
  - *Áreas de diseño: ubicación de los datos, paralelización automática, control de flujo de datos y recuperación de datos. Usa multihilos.*
- Crace (Universidad de Tokio-84)
  - *Es una máquina de BD paralela que contiene módulos para: procesamiento, memoria, disco y control. Procesamiento basado en el flujo continuo (streams) de datos. Interconexión de los módulos con 2 buses: el anillo de procesamiento y el de montaje. Usa hashing dinámico y mezcla-ordenamiento con pipeline.*

UNIVERSIDAD DE LOS ANDES

SGBDRP

- SDC (Universidad de Tokio-90)
  - *Basado en nada se comparte.*
  - *Cada módulo de procesamiento es todo se comparte interconectados con la red omega.*
  - *Usa repartición en cubetas y ubicación dinámica de las mismas según su tamaño y la carga de los módulos de procesamiento.*
- Volcano (Universidad de Colorado-90)
  - *Contiene un operador de intercambio que paraleliza los demás operadores.*
  - *No soporta paralelismo entre consultas.*

UNIVERSIDAD DE LOS ANDES

Formas de paralelismo

- Paralelismo intra-operadores:
  - ✓ *Varios procesadores efectúan la misma operación sobre diferentes subconjuntos de datos compartidos.*
- Paralelismo entre-operadores:
  - ✓ *Vertical: entuba los datos entre los procesadores.*
  - ✓ *Horizontal: Diferentes procesadores ejecutan diferentes subárboles del árbol de consulta.*

UNIVERSIDAD DE LOS ANDES

Procesamiento de consultas

- Paralelismo intra-consulta:
  - ✓ *Partes independientes de una consulta se ejecutan en paralelo.*
  - ✓ *Se implementa en forma limitada dividiendo las relaciones en disco y ejecutando el procesamiento de cada partición en paralelo (DBC).*
- Paralelismo entre-consultas:
  - ✓ *Varias consultas se ejecutan independientemente y en paralelo.*
  - ✓ *Se implementa fácilmente en el modelo de todo se comparte (DBC).*
- División de los datos (desagrupamiento):
  - ✓ *Lectura/escritura paralela de datos en los discos.*
  - ✓ *1 relación en varios D.*

UNIVERSIDAD DE LOS ANDES

Procesamiento de consultas

- *Round-robin:*
  - ✓ *Las tuplas se colocan en disco en forma circular, obteniendo aprox. el mismo número de tuplas en cada disco.*
  - ✓ *Esto balancea la carga de E/S para las consultas que requieren un barrido secuencial de la relación (Bubba).*
- *Hashing: (bueno en consultas exactas)*
  - ✓ *Almacena las tuplas en los discos en base a una función hashing que permite dirigir la recuperación hacia uno de ellos (Bubba).*

UNIVERSIDAD DE LOS ANDES

Procesamiento de consultas

- *División por rango:*
  - ✓ *Las tuplas se colocan en los discos según una función de rango, normalmente en base a la condición del WHERE.*
  - ✓ *Problema: desbalanceo de la carga entre los discos debido a la distribución de los datos por la condición.*
  - ✓ *Ejm: departamentos entre 1 y 10 en el disco 1, entre 11 y 20 en el disco 2, etc. (Bubba).*
- *División híbrida por rango (Gamma):*
  - ✓ *Desagrupa una relación según: los requerimientos de dispositivos de las consultas que acceden a la relación, las capacidades de procesamiento de los procesadores y el overhead obtenido si se usa un procesador adicional para ejecutar la consulta.*

UNIVERSIDAD DE LOS ANDES

Algoritmos paralelos de productos

- Equiproducto:
  - ✓ Paralelizar estos algoritmos dividiendo el problema y los datos en piezas independientes para procesarlas en paralelo.
  - ✓ *Método de los bloques anidados:*
    - *Variación de los lazos anidados que toma la división de una relación en páginas para reducir el costo de E/S.*
    - *Es eficiente si hay suficiente memoria principal para almacenar la relación más pequeña.*
    - *Para el modelo de nada se comparte, ambas R y S están desagrupadas y diseminadas en los discos.*
    - *Cada procesador difunde su porción de R a los otros procesadores, por lo que cada uno tiene R completa y puede aplicar el método en cada nodo.*

UNIVERSIDAD DE LOS ANDES

Algoritmos paralelos de productos

- *Método de ordenamiento y mezcla:*
  - *Ambas R y S se ordenan por el atributo del equiproducto utilizando un algoritmo paralelo y luego se dividen diseminándolas en los nodos usando condiciones de rango, para la fase de mezcla.*
  - *Necesita menos E/S que el anterior.*
- *Método de producto hash:*
  - *Se dividen R y S en subconjuntos disjuntos usando una función hash sobre el atributo del producto y se almacenan las porciones en cubetas diferentes, por lo que las tuplas con el mismo valor del atributo están en la misma cubeta y pueden ser concatenadas las de R y S fácilmente.*
  - *Versiones: producto hash simple, producto hash de Grace y el híbrido.*
  - *Con multiprocesadores se tiene que se asigna una cubeta por procesador y todos hacen el trabajo en paralelo*

UNIVERSIDAD DE LOS ANDES

Algoritmos paralelos de productos

- Datos sesgados: Cuando unos valores de un atributo ocurren más frecuentemente que otros, haciendo que unos procesadores tengan más tuplas que otros (desbalance de datos) y un desbalance de carga, que hace que un procesador trabaje más que los otros y se convierta en un cuello de botella (carga de trabajo desbalanceada).
- *El problema es diseñar un algoritmo inmune a los datos sesgados.*
- *En el modelo nada se comparte: Opciones:*
  - ✓ *Modificación del método de ordenamiento y mezcla para tratar los datos sesgados [Wolf-90]. Se usa un algoritmo de optimización que toma la salida de la fase de ordenamiento y determina como será dividido en múltiples tareas y como esas tareas deben ser asignadas a los procesadores para que la carga este balanceada.*
  - ✓ *Modificación del método de producto hash donde las cubetas se dividen en fragmentos y ellas son diseminadas en los procesado-res para obtener una carga balanceada.*

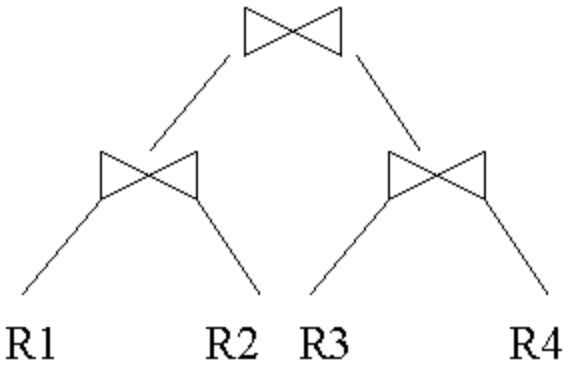
UNIVERSIDAD DE LOS ANDES

Algoritmos paralelos de productos

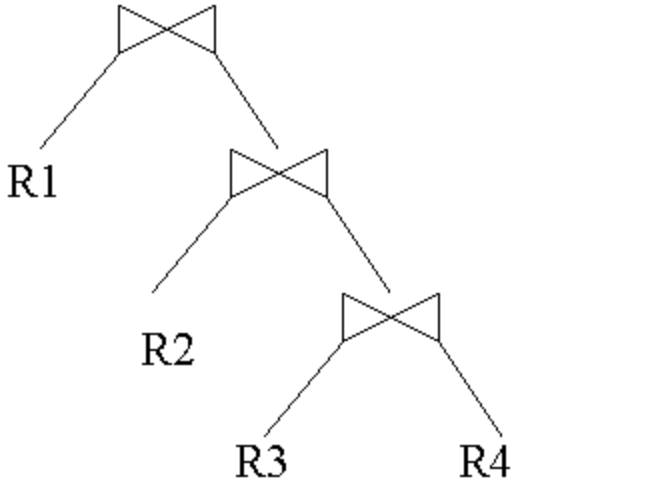
- *En el modelo todo se comparte:*
  - ✓ *Se usa otra modificación del método de producto hash.*
- *Si se tiene una BD grande en un multiprocesador en hipercubo, el mejor algoritmo es el producto hash cubo adaptativo que es relativamente inmune a la distribución de los datos*
- *Ejecución de consultas multiproducto:*
  - ✓ *Si hay varias operaciones de producto en una consulta.*
  - ✓ *Si el árbol de la consulta es:*
    - o *Balanceado: los productos de cada rama se pueden realizar en paralelo*
    - o *Desbalanceado a la derecha: los productos se hacen en pipeline segmentando la rama derecha del árbol.*
- *Con un ambiente de BD paralelo el número de planes de ejecución de una consulta es mayor (puede ser exponencial).*

UNIVERSIDAD
DE LOS ANDES

árbol de la consulta



Balanceado



Desbalanceado a la derecha

UNIVERSIDAD DE LOS ANDES

Paralelismo en DB2 v.3

- Provee soporte para realizar en paralelo las operaciones de E/S creando tablas divididas y esparcidas en disco, para consultas SQL estáticas y dinámicas, locales o remotas.
- Se usa para productos de una sola tabla o multitabla.
- DB2 estima si debe usarlo o no, dependiendo si el tiempo de E/S > tiempo de UCP.
- El manejador del buffer soporta E/S paralela para condicionar el tamaño de la página y para la disponibilidad del pool de buffers