

Maestría en Ciencia de Datos

Minería de Datos

Tarea 02: Paralelismo

Profesor:

Dr. Ángel Fernando Kuri Morales

Alumna:

Gabriela Flores Bracamontes

Clave única:

160124

México, D.F. 24 de agosto de 2015.

1. Datos de mi equipo de cómputo

- Laptop Lenovo – ThinkPad W540
- Sistema Operativo Centos 7
- Memoria - 32 GB RAM
- Intel® Core™ i7-4900MQ (hasta 2,8 GHz, L3 de 8 MB, 1600 MHz FSB) de 4 núcleos
- NVIDIA Quadro K2100M de 2 GB
- Disco Duro 250GB - Estado Sólido

2. Paralelismo

El procesamiento paralelo consiste en acelerar la ejecución de un programa mediante la descomposición en fragmentos que puedan ejecutarse de forma paralela, cada uno en una unidad de proceso diferente. Una forma de clasificar la computación, es a través de dos eras de desarrollo: Era de computación secuencial y Era de computación paralela.

La razón principal para crear y utilizar computación paralela es que el paralelismo es una de las mejores formas de salvar el problema del cuello de botella que significa la velocidad de un único procesador.

Las aplicaciones que se benefician de una aceleración más significativa son aquellas que describen procesos intrínsecamente paralelos, las simulaciones de modelos moleculares, climáticos o económicos tienen toda una amplia componente paralela, como los sistemas que representan. el hardware de la máquina entra en juego ya que es preciso maximizar la relación entre el tiempo de cálculo útil y el perdido en el paso de mensajes, parámetros que dependen de la capacidad de proceso de las CPUs y de la velocidad de la red de comunicaciones.

Hay 2 formas básicas de obtener partes independientes en un programa paralelo: descomposición funcional o descomposición de datos, que describiremos a continuación.

Descomposición de datos. Un ejemplo de aplicación completamente paralelizable es el cálculo del área bajo una curva por integración numérica, basta con dividir el intervalo de integración entre todos los procesadores disponibles y que cada uno resuelva su fragmento sin preocuparse de qué hacen los demás, al final, los resultados parciales se recolectan y se suman convenientemente.

Con n procesadores es posible resolver el problema n veces más rápido que haciendo uso de uno sólo (salvo por el mínimo retraso que supone el reparto de trabajo inicial y la recolección de datos final), consiguiendo una aceleración lineal con el número de procesadores. Si las condiciones son muy favorables es incluso posible alcanzar la aceleración superlineal, consistente en que el programa se ejecuta aún más rápido que en régimen línea, la aparente paradoja se da debido a que cada procesador cuenta con su propia memoria ordinaria y caché, que pueden ser usadas de forma más eficiente con un subconjunto de datos, de hecho, es posible que el problema no se pueda resolver en un único procesador pero sí sobre un conjunto de ordenadores debidamente configurados, simplemente por cuestión de tamaño de los datos.

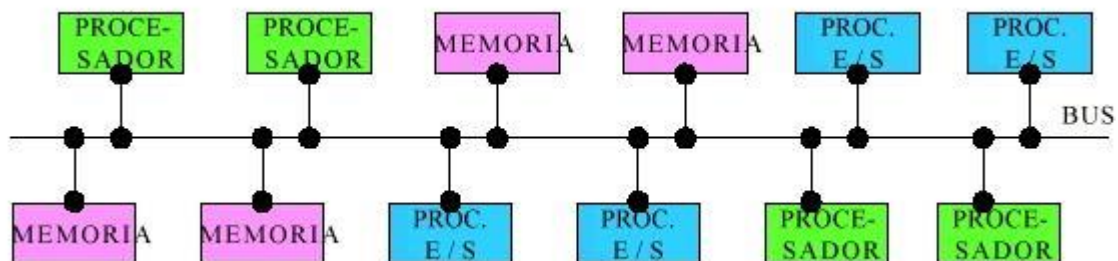
Descomposición funcional. Un modelo computacional se basa por empezar, en que una aplicación consiste en varias tareas, cada tarea es responsable de una parte de la carga de procesamiento de la aplicación en general y a su vez, cada tarea realiza una operación independiente de las otras tareas. Los algoritmos de cada tarea son diferentes, este modelo se denomina descomposición funcional y se puede aprovechar las características particulares de cada tipo de tarea para ejecutarlas en la máquina que sea más conveniente para tal efecto.

EVOLUCIÓN

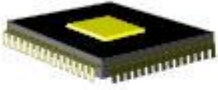



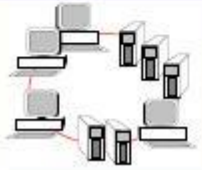

Durante años se han encontrado dificultades al momento de analizar sistemas de gran tamaño; si bien en el caso de los sistemas de potencia éste se ha visto favorecido por la descripción de problemas mediante matrices, y por la utilización de computadores digitales para su manipulación, la necesidad de lograr un equilibrio en la cantidad de información a procesar y su calidad continúa siendo evidente.

El procesamiento paralelo ha permitido sobrellevar algunas de estas dificultades, particularmente en lo que respecta a la velocidad de procesamiento; siempre que la arquitectura del computador sea la apropiada.

Los sistemas paralelos mejoran la velocidad de procesamiento y de E/S mediante la utilización de CPU y discos en paralelo. Cada vez son mas comunes computadoras paralelas, lo que hace que cada vez sea mas importante el estudio de los sistemas paralelos de bases de datos.



En el proceso paralelo se realizan muchas operaciones simultáneamente, mientras que en el procesamiento secuencial los distintos pasos computacionales han de ejecutarse en serie, la mayoría de las máquinas de gama alta ofrecen un cierto grado de paralelismo de grano grueso: son comunes las máquinas con dos o cuatro procesadores. Las computadoras masivamente paralelas se distinguen de las máquinas de grano grueso porque son capaces de soportar un grado de paralelismo mucho mayor.

nivel		técnicas de implementación	
estructuras paralelas		paralelismo a nivel de procesador	 <ul style="list-style-type: none"> - segmentación - división funcional - procesadores vectoriales
		paralelismo en multiprocesadores	 <ul style="list-style-type: none"> - memoria compartida - memoria distribuida
		paralelismo en multicomputadores	 <ul style="list-style-type: none"> - <i>clusters</i> - sistemas distribuidos

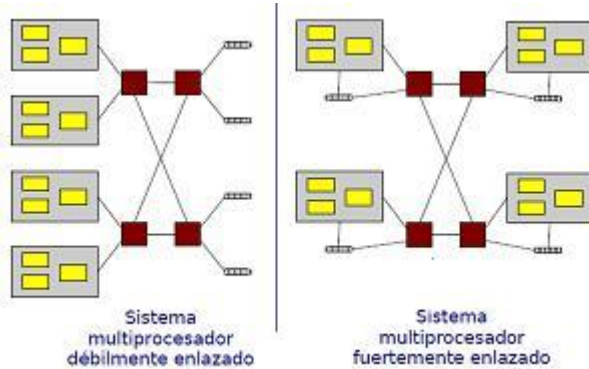
Ya se encuentran en el mercado computadoras paralelas con cientos de CPU's y discos, para medir el rendimiento de los sistemas de bases de datos existen dos medidas principales, la primera es la productividad: número de tareas que pueden completarse en un intervalo de tiempo determinado, la segunda es el tiempo de respuesta: cantidad de tiempo que necesita para complementar una única tarea a partir del momento en que se envíe.

Un sistema que procese transacciones puede mejorar el tiempo de respuesta, así como la productividad, realizando en paralelo las distintas subtareas de cada transacción, actualmente se han propuesto soluciones basadas en procesamiento paralelo para una gran cantidad de los análisis asociados a los sistemas; entre éstos se encuentran el flujo de potencia, la simulación dinámica, incluyendo la de transitorios electromagnéticos; el análisis de la estabilidad transitoria mediante funciones de energía, el de la estabilidad a perturbación pequeña y a la planificación de sistemas eléctricos de potencia.

El procesamiento paralelo es particularmente atractivo si es posible descomponer el sistema en subsistemas acoplados débilmente, pero cada uno con sus variables acopladas fuertemente.

CARACTERÍSTICAS

El uso de varios procesadores está motivado por consideraciones relativas a las prestaciones y/o a la fiabilidad, podemos clasificar dichos sistemas como sigue:



- Multiprocesadores débilmente acoplados - Consisten en un conjunto de sistemas relativamente autónomos, en los que cada CPU dispone de su propia memoria principal y sus canales de E/S. En este contexto se utiliza frecuentemente el término multicomputador.
- Procesadores de Uso Específico - Tales como un procesador de E/S. En este caso, hay un maestro, una CPU de uso general, y los procesadores de uso específico están controlados por la CPU maestra a la que proporcionan ciertos servicios.
- Multiprocesadores fuertemente acoplados - Constituidos por un conjunto de procesadores que comparten una memoria principal común y están bajo el control de un mismo sistema operativo.
- Procesadores paralelos - Multiprocesadores fuertemente acoplados que pueden cooperar en la ejecución en paralelo de una tarea o un trabajo.

El procesamiento en paralelo se basa principalmente en Multiprocesadores fuertemente acoplados que cooperan para la realización de los procesos, aquí sus características.

- Posee dos o más procesadores de uso general similares y de capacidades comparables.
- Todos los procesadores comparten el acceso a una memoria global.
- También pueden utilizarse algunas memorias locales (privadas como la cache).
- Todos los procesadores comparten el acceso a los dispositivos de E/S, bien a través de los mismos canales bien a través de canales distintos que proporcionan caminos de acceso a los mismos dispositivos.
- El sistema está controlado por un sistema operativo integrado que permite la interacción entre los procesadores y sus programas en los niveles de trabajo, tarea, fichero, y datos elementales.

La ganancia de velocidad y la ampliabilidad son dos aspectos importantes en el estudio del paralelismo, la ganancia de velocidad se refiere a la ejecución en menor tiempo de una tarea dada mediante el incremento del grado de paralelismo, la ampliabilidad se refiere al manejo de transacciones más largas mediante el incremento del grado de paralelismo.

Considérese un sistema paralelo con un cierto número de procesadores y discos que está ejecutando una aplicación de base de datos, supóngase ahora que se incrementa el tamaño del sistema añadiéndole más procesadores, discos y otros componentes. La ampliabilidad está relacionada con la capacidad para procesar más largas e el mismo tiempo mediante el incremento de los recursos del sistema.



VENTAJAS Y DESVENTAJAS

Existen algunos factores que trabajan en contra de la eficiencia del paralelismo y pueden atenuar tanto la ganancia de velocidad como la ampliabilidad.

- **Costes de inicio:** en una operación paralela compuesta por miles de proceso, el tiempo de inicio puede llegar ser mucho mayor que le tiempo real de procesamiento, lo que influye negativamente en la ganancia de velocidad.
- **Interferencia:** como lo procesos que se ejecutan en un proceso paralelo acceden con frecuencia a recursos compartidos, pueden sufrir un cierto retardo como consecuencia de la interferencia de cada nuevo proceso en la competencia, este fenómeno afecta tanto la ganancia de velocidad como la ampliabilidad.
- **Sesgo:** normalmente es difícil dividir una tarea en partes exactamente iguales, entonces se dice que la forma de distribución de los tamaños es sesgada.

El procesamiento paralelo implica una serie de dificultades a nivel programación de software, es difícil lograr una optimización en el aprovechamiento de los recursos de todas las CPU con el que se esté trabajando sin que se formen cuello de botella. En muchas de las ocasiones no es posible el trabajar con equipos multiprocesadores dado el elevado costo que este representa, así que solo se dedica a ciertas áreas de investigación especializada o proyectos gubernamentales o empresariales.

Ventajas del Procesamiento en Paralelo.

El procesamiento en paralelo ejecuta procesos en donde cada procesador se encarga de uno u otro y aceleran de esta forma el cálculo.