¿Qué hay de nuevo en las CPUs desde los 80s?

MISCELÁNEA

Los chips adquirieron registros más amplios y pueden ocuparse de más memoria. Los CPUs pasaron de ser de 8 bits a ser de 64 bits.

SERIALIZACIÓN

Fue implementada como solución para ordenar con precisión instrucciones y que cada una se ejecute en el tiempo que debe ejecutarse.

CAMBIOS DE CONTEXTO/ LLAMADAS AL SISTEMA

Los cambios de contexto en el almacenamiento caché de los núcleos modernos es demasiado costoso y por ende las llamadas al sistema son costosas. Por estas razones es que las personas han optado por usar versiones por lotes de llamadas al sistema para código de alto rendimiento.

SIMD

SIMD (Single Instruction, Multiple Data) otorga ventajas al utilizar este tipo de instrucciones ya que se utilizan para acelerar los procesos. Estas instrucciones realizan una misma operación para un grupo grande de datos, reconoce patrones y de ahí optimiza ese proceso.

ADMINISTRACIÓN DE ENERGÍA

Actualmente hay muchas características sofisticadas para administrar la energía de las CPUs . Se ha demostrado que micro-optimizaciones específicas pueden beneficiar el consumo de energía.

MEMORIA

CACHÉ

Los procesadores aumentaron su velocidad, pero las memorias no, por lo que la solución para esto fue agregar el almacenamiento caché el cual otorga acceso más rápido a datos de uso frecuente.

CONCURRENCIA

Las cargas y almacenamientos a una misma ubicación trajeron consigo pérdida o traslape de información y el uso de serialización para todo sería un proceso muy lento, por lo que se implementó el uso del prefijo 'lock' para volver atómicas las instrucciones.

COMBINACIÓN DE ESCRITURA

La memoria combinada de escritura (WC) es una especie de memoria no almacenable en caché (UC) que eventualmente es consistente porque las escrituras en algún momento llegan a la memoria, pero pueden almacenarse en el búfer internamente.

NUMA

Non-uniform memory access se utiliza en el multiprocesamiento donde la memoria se accede en posiciones relativas de otro proceso o memoria compartida entre procesos.

GPU / GPGPU

La arquitectura paralela de la GPU podría funcionar para utilizarse en un proceso que tuviera varios subprocesos simultáneos, pero el código para esto tenía que usar el API de los gráficos tradicionales los cuales estaban muy limitados. Nvidia y ATI observando esto lanzaron frameworks que permitían acceder a mayor parte del hardware, con esto actualmente las GPU son utilizadas ampliamente para la computación de alto rendimiento junto con las CPU.

INTERFACES

Las GPU modernas deben tener una CPU para copiar datos hacia y desde la memoria de la CPU y la GPU, y para iniciar y codificar en la GPU. En el rendimiento más alto, un bus PCIe 3.0 con 16 carriles puede alcanzar velocidades de alrededor de 13-14 GB/s. A medida que las GPU se vuelven más potentes, el bus PCIe se convierte cada vez más en un cuello de botella.

PROCESADORES

Un GPU contiene uno o varios multiprocesadores de transmisión (SM) y cada uno de ellos contiene más de 100 unidades de punto flotante o mejor conocidos como núcleos en los GPUs. Cada núcleo tiene una velocidad de reloj de aproximadamente 800 MHz.

MEMORIA

La memoria del GPU se divide en 3: memoria global (GDDR), mem-oria compartida y registros. Tanto la global como la compartida tienen reglas estrictas para acceder a ellas y para alcanzar el mejor rendimiento los accesos a la memoria deben fusionarse completamente entre subprocesos dentro del mismo grupo de subprocesos.

THREADING MODEL

Los subprocesos del GPU se ejecutan en forma SIMT (Single Instruction Multiple Thread), y cada subproceso se ejecuta en un grupo de tamaño predefinido del hardware. Cada subproceso en ese grupo debe trabajar en la misma instrucción al mismo tiempo y si algún subproceso necesita tomar una ruta diferente de código, todos los subprocesos que no forman parte de la rama suspenden la ejecución hasta que se complete.

GLOBAL

Tiene un tamaño entre 2 y 12 GB con un rendimiento de 300 a 400 GB/s. Todos los subprocesos de todos los SM del procesador pueden acceder a ella. Es el tipo de memoria más lento de la tarjeta.

COMPARTIDA

La comparten todos los subprocesos dentro del mismo SM. Es el doble de rápida que la memoria global, pero no es accesible entre subprocesos en diferentes SM.

REGISTROS

Son similares a los registros de la CPU en el sentido de que son la forma más rápida de acceder a los datos en una GPU, pero son locales a un subproceso y los datos no son visibles para ninguno otro subproceso.

Rojo López Luis Felipe