UNIVERSIDAD DE GRANADA

Master: Ingeniería de Software

Asignatura: Computación de propósito general en unidades de procesamiento grafico.

Apellidos: Gil

Nombre: Jonas

Declaración personal de no plagio
1. Tengo conocimiento de que plagiar supone usar el trabajo de otro y presentarlo como propio, y de que constituye una infracción de los derechos de propiedad intelectual. **2. Declaro que lo que aquí presento es fruto mi propio trabajo.** **3. No he permitido, y no permitiré, que nadie copie mi trabajo con la intención de hacerlo pasar como su propio trabajo.**

Tarea a realizar:

Resumen de la lectura que se adjunta (capítulo 1 del libro "GPU Parallel Program Development Using CUDA" de Tolga Soyata) que incluya una reflexión crítica al final del documento sobre los contenidos del mismo y su repercusión.

El texto proporciona una visión histórica y crítica del papel de la supercomputación en la evolución de la tecnología informática, con un enfoque particular en la relevancia de las GPUs y la tecnología CUDA se destaca cómo los avances en supercomputación, como el uso de GPUs en máquinas de alto rendimiento, eventualmente se traducen en avances en la informática de escritorio.

La supercomputación impulsa muchas de las tecnologías presentes en los procesadores modernos, ya que la necesidad de procesar conjuntos de datos cada vez más grandes ha llevado a la producción de computadoras más rápidas, este impulso tecnológico ha llevado a la convergencia entre el cómputo de alto rendimiento y las computadoras personales.

Por otro lado se destaca el cambio hacia una computación heterogénea, combinando CPU y GPU para lograr un mejor rendimiento como es el caso de proyectos de computación distribuida como BOINC y Folding@Home, que permiten la contribución de personas comunes a proyectos científicos.

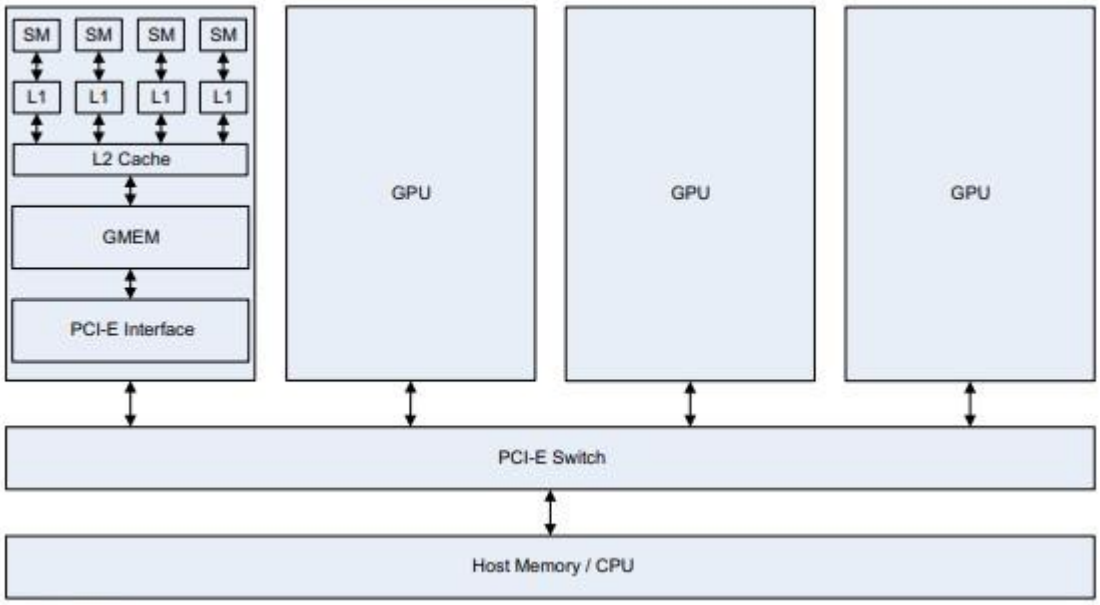

Imagen 1: Combinación de GPU y CPU

Según el auto la arquitectura de Von Neumann subyace en la mayoría de los procesadores modernos y se exploran los desafíos asociados con la limitación de la velocidad de procesamiento por la velocidad de la memoria. Se discute el uso de la caché para mitigar este problema y cómo el tamaño de la caché influye en el rendimiento y el costo de los procesadores.

Se menciona brevemente la historia de la supercomputación, donde se destacan los logros de Seymour Cray y el diseño revolucionario de la Connection Machine, que utilizaba una gran cantidad de procesadores en paralelo para realizar tareas computacionales y e discuten las ventajas y desafíos de este enfoque altamente paralelo en comparación con las arquitecturas de CPU convencionales.

El texto ofrece una panorámica crítica y reflexiva sobre la evolución de la supercomputación y su impacto en la tecnología informática actual, resaltando la importancia de comprender y aprovechar eficazmente las tecnologías emergentes, como las GPUs y la computación heterogénea, en el desarrollo de software y hardware avanzados.

El informe presenta una revisión de los primeros días de la programación de GPGPU (General-Purpose Graphics Processing Unit), destacando el papel evolutivo de las GPUs y la tecnología CUDA. Comienza explicando la función básica de las GPUs en la representación de imágenes y la evolución de los efectos fotorealistas, se señala la importancia del desarrollo de sombreadores programables como el primer paso hacia la programación de propósito general en las unidades de procesamiento gráfico.

A pesar de estos avances, se destaca que la programación GPGPU inicial enfrentó desafíos debido a la falta de un lenguaje de programación generalizado y a la complejidad de las herramientas disponibles y que la introducción de CUDA por parte de NVIDIA en 2007 marcó un cambio significativo al proporcionar una interfaz de programación más accesible y eficiente, lo que llevó a un aumento en la adopción de la programación GPU.

Se aborada el cambio en la estrategia de desarrollo de procesadores, pasando de aumentar la velocidad de reloj a agregar más núcleos, resaltándose la necesidad de adaptarse a la programación multicore y la importancia de aprovechar al máximo el potencial de paralelización.

Se discute el crecimiento en la potencia computacional de las GPUs en comparación con las CPUs, destacando el aumento en el rendimiento de las GPUs a través de la introducción de arquitecturas masivamente paralelas, se enfatiza el impacto de CUDA en la transformación de la computación de alto rendimiento y se menciona la creciente disponibilidad de aplicaciones compatibles con CUDA.

Finalmente, se proporciona información sobre la arquitectura de las GPUs NVIDIA y las consideraciones de hardware al construir sistemas de computación acelerada. Se menciona la disponibilidad de racks de Tesla y la viabilidad de construir clústeres GPU utilizando componentes estándar de PC.

El informe examina alternativas a CUDA, centrándose en OpenCL, DirectCompute y alternativas de CPU, destacando que OpenCL es una norma abierta compatible con múltiples fabricantes de GPU, permitiendo el uso de dispositivos de cálculo como GPU,

CPU u otros dispositivos especializados. Aunque similar a CUDA, OpenCL es más complejo de usar y requiere realizar más trabajo manual por parte del programador.

DirectCompute, respaldado por Microsoft, es una alternativa propietaria vinculada al sistema operativo Windows y la API DirectX 11., Aunque ofrece una transición relativamente sencilla desde CUDA para desarrolladores familiarizados con DirectX, su alcance está limitado a entornos Windows.

Para desarrolladores que prefieren trabajar en entornos CPU, se mencionan alternativas como MPI, OpenMP y pthreads para Linux, y el modelo de subprocesos de Windows y OpenMP para Windows, Estas opciones permiten la programación paralela dentro de nodos o sistemas de computadora, pero su eficacia puede variar según la arquitectura subyacente del CPU.

Además, se mencionan bibliotecas como ZeroMQ y Hadoop, junto con directivas de compilador como OpenACC, que facilitan la programación paralela, Aunque CUDA sigue siendo la opción más popular debido a su facilidad de uso y soporte, entender estas alternativas es importante para los desarrolladores interesados en optimizar el rendimiento y la escalabilidad de sus aplicaciones.

En conclusión, la evolución de la supercomputación, impulsada por las GPUs y la tecnología CUDA, ha sido fundamental en la transformación de la informática, la convergencia entre el cómputo de alto rendimiento y las computadoras personales ha llevado a avances significativos por lo que es crucial adaptarse y aprovechar tecnologías emergentes para optimizar el rendimiento y la escalabilidad de las aplicaciones, la programación GPGPU ha enfrentado desafíos iniciales, pero la introducción de CUDA ha marcado un cambio notable, facilitando una interfaz de programación más accesible y eficiente, además, el cambio en la estrategia de desarrollo de procesadores, pasando de la velocidad de reloj a agregar más núcleos, resalta la importancia de la programación multicore y la paralelización, por lo que el impacto de CUDA en la transformación de la computación de alto rendimiento es innegable, con un aumento en la disponibilidad de aplicaciones compatibles y aunque CUDA sigue siendo popular, también existen otras alternativas como OpenCL y DirectCompute, así como otras opciones de CPU, para optimizar el rendimiento y la escalabilidad de las aplicaciones en diferentes entornos.

-