Intel Xeon Phi

Miguel Angel Vidal, Yesid Leandro Marquez, Cesar Augusto Sarmiento
Ingeniería de sistemas y computación, Universidad tecnológica de Pereira, Pereira, Colombia
Correo-e: miguel031926@gmail.com, marquez2388@utp.edu.co, cesarmiento17@utp.edu.co

***Resumen*—** Xeon phi es una arquitectura de alto nivel diseñada para tecnología que implica un alto rendimiento, el cual tiene más de 50 núcleos para el caso del artículo base pertenece a la serie 5100 y posee 60 núcleos, conectados por una interconexión bidireccional de alto rendimiento, 16 canales de memoria, cuando se trabaja como acelerador, Phi puede ser Conectado a un host (es decir, un dispositivo que lo gestiona) a través de Una interfaz de sistema PCI Express (PCIe) Similar a un Acelerador GPU.

***Palabras clave*—** Ancho de banda, Arquitectura, Benchmarking, Chip, Ciclos, Compilador, CPU, GPU, Hardware, Hilos, Interconexión, Iteración, Kernel, Latencia, Memoria caché, Núcleos, Prefetching, Software, Supercomputadoras, Xeon Phi

I. INTRODUCCIÓN

En el siguiente artículo hablaremos de la arquitectura de Xeon phi el cual es hecho por Intel es un procesador de gran rendimiento por lo tanto no es muy común escuchar o leer sobre estos procesadores pero gracias a algunas investigaciones y a un artículo el cual fue la base para la realización de este mismo, en dicho artículo "An study empirical of Intel Xeon phi" realizan varias pruebas al procesador, para demostrar que tan rápido es y que tan bueno es para programar sobre él, , también podremos ver algunas diferencias con la actualidad del Xeon phi el cual tuvimos conocimiento este año realizó una nueva serie estando así a un mayor nivel y estando así bien ubicado entre las mejores arquitecturas.

Todo esto a partir de querer mostrar la superioridad de Intel Xeon phi y pues mostrar los avances que ha tenido esta buena empresa con sus procesadores colaborando así al avance tecnológico para las supercomputadoras.

II. CONTENIDO

Novedades de esta arquitectura:
Los núcleos de procesamiento vectorial, El chip on-chip Memoria, La memoria fuera de chip, La interconexión de anillo,
La conexión pcie.

El Xeon phi posee una interconexión de anillo bidireccional rápida. Todas Las entidades conectadas utilizan el anillo para fines de comunicación, usando controladores especiales llamados paradas de anillo para insertar solicitudes Y recibir respuestas en el anillo. Cada núcleo contiene una unidad de vector de 512 bits de ancho (VPU) con archivos de registro vectorial (32 registros por contexto de hilo). Cada núcleo tiene un caché de datos LK de 32KB, una caché de instrucción LK de 32KB y un caché L2 unificado de 512KB. [1]

Los núcleos de procesamiento vectorial: por cada Ciclo (1TFlops en total) - se puede lograr al usar 240 Hilos y la instrucción multiplica-agrega esto es dos veces mayor al rendimiento mul también hay dos observaciones más dentro de la arquitectura cuando se utilizan 60 (Un hilo por núcleo), el rendimiento de la instrucción es bajo en comparación con los casos cuando se utilizan 120 o 240 hilos. Esta se debe al hecho de que no es posible emitir instrucciones desde el mismo contexto de hilo en back-to-back ciclos así, los programadores necesitan correr al menos dos hilos en cada núcleo para poder utilizar completamente los recursos de hardware quizás para los usuarios normales el Xeon phi pueda ser usado con 60 hilos por núcleo, pero estos procesadores, esta arquitectura está diseñada para supercomputadoras que no van a ser usadas por usuarios normales sino para programadores debido a su gran rendimiento ya que para los usuarios normales no es necesario tanta capacidad de rendimiento.

Cuando un hilo utiliza sólo una secuencia de instrucciones, tenemos que utilizar 4 hilos por núcleo (240 hilos en total) para lograr el rendimiento máximo de la instrucción. Esto se debe a que la latencia de una instrucción aritmética es de 4 ciclos, [1]

esto demuestra lo que decíamos anteriormente que para optimizar el rendimiento del Xeon phi es necesario la mayor cantidad de hilos posibles por núcleo, para que al momento de la programación sea mucho más rápido.

Continuaremos hablando sobre algunas especificaciones del Xeon phi para así entender bien el tema sobre su rendimiento.

1- Latencia de memoria: Daniel Molka , propuso un Enfoque para cuantificar la caché coherencia efectos en esta aplicación adaptaron esta aplicación a xeon phi . Esencialmente, la Aplicación atraviesa una matriz A de tamaño S ejecutando k = A [k] En un bucle totalmente desenrollado. La matriz se inicializa con un paso Es decir, A [k] = (k + stride)% S. Al medir el tiempo de ejecución Del recorrido, podemos obtener fácilmente una estimación del promedio Tiempo de ejecución para una iteración. Vemos que el Xeon Phi tiene dos niveles de caché de datos (L1 y L2). Los datos L1 Caché es de 32 KB, mientras que los cachés de datos L2 deben ser más pequeños Que 512KB. Además, la latencia de acceso de L1 y L2 caché de datos es de alrededor de 2,87 ns (3 ciclos) y 22,98 ns (24 ciclos), respectivamente. Con un paso de 64 bytes, Xeon Phi Toma 287.51? 291,18 ns (302 - 306 ciclos) para terminar un Acceso a datos en la memoria principal (cuando el conjunto de datos es de 512KB). [1]

2- Latencia del caché remoto: Por lo tanto, en esta sección, nos centramos en medir Latencia de caché. Enfoque basado por Daniel Molka Cada medida, se utilizan dos hilos (T0, T1), con T0 Fijado al Núcleo 0 y T1 fijado en otro núcleo (Núcleo X). La medida de latencia siempre se ejecuta en Core 0, transfiriendo Un número predefinido de líneas de caché de Core X a Core 0), se logra el mejor rendimiento Cuando cada núcleo mantiene los datos en su propia caché local, y Evita la memoria caché remota y los accesos de memoria como sea posible. Con respecto a la capacidad de caché, el Xeon Phi Utiliza una LLC privada (cache de último nivel) de tamaño 512KB Específicamente, si ningún núcleo comparte ningún dato o código, la El tamaño total L2 del chip es de 30 MB (con 60 núcleos). Mientras, Si cada núcleo comparte exactamente los mismos datos y código en perfecto Sincronización, entonces el tamaño L2 total efectivo del chip Es 512 KB Para ahorrar el escaso recurso on-chip. [1]

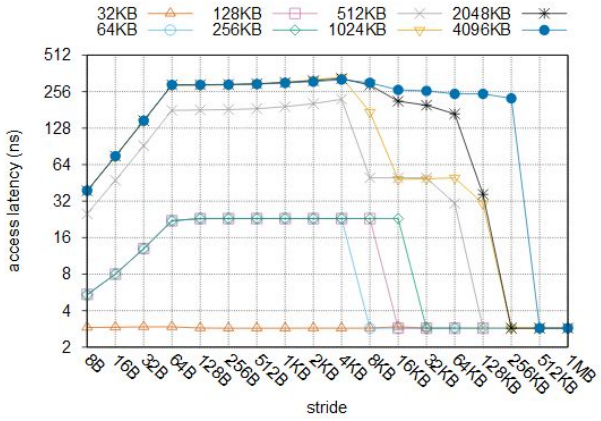

Fig 1. Rendimiento de la latencia

3- Ancho de banda de la memoria

Daniel Molka. Presenta una solución para medir el ancho de banda De forma similar a la de la medición de la latencia Su micro benchmark requiere compilador Optimizaciones a desactivar (es decir, el código debe ser compilado Con la opción -O0), desactivando así el prefetching de software En Xeon Phi. Como resultado, esta medida subestimará Ancho de banda Lmbench intenta medir la capacidad del sistema Para transferir datos entre procesador, caché, memoria, disco y red caso específico. Incluye una revisión sistemática comentada de la literatura sobre casos análogos.

4- Ancho de Banda de Memoria Sin Chip: Vea que el ancho de banda máximo para leer y escribir Está muy por debajo del máximo teórico de 320 GB / s. Además, El ancho de banda de memoria de lectura y escritura aumenta El número de hilos, lo que ocurre porque al utilizar Más hilos, el ancho de banda de lectura alcanza un máximo de 164 GB / s, alcanzable con Usando 60 hilos o más (fijando al menos un hilo a un núcleo). Sin embargo, podemos obtener el ancho de banda de escritura máximo (76 GB / s,) sólo cuando se utilizan 240 hilos. En General, el ancho de banda de escritura es alrededor de la mitad de la lectura Ancho de banda Esto sucede porque Xeon Phi implementa una Escribir-asignar la política de caché y el contenido original tiene que ser Cargado en caches antes de que lo sobrescribamos por completo.

5- Ancho de banda agregado: Interconexión de anillo En Xeon Phi, los núcleos y controladores de memoria son

inter-Conectado en un anillo bidireccional. Cuando varios subprocesos Solicitando datos simultáneamente, componentes compartidos como el Ring stop o DTD pueden convertirse en cuellos de botella de rendimiento. de la memoria de la En-Viruta: El disponible El ancho de banda de la memoria en el chip siempre es esencial en el rendimiento Afinación y análisis. Así que, ¿cuán grande es la memoria en el chip Nuestros resultados Muestran que el rendimiento de acceso L1 (lectura o escritura) es de 64 bytes Por ciclo. Por lo tanto, el ancho de banda L1 agregado es 4032 GB / s para Leer o escribir. Luego medimos el máximo alcanzado Ancho de banda desde el punto de vista de los programadores para scale1 (O [i] = a × A [i]), scale2 (O [i] = a × O [i]), saxpy1 (O [i] = a × A [i] + B [i]), y saxpy2 (O [i] = a × A [i] + O [i] Operaciones. Para evitar gastos generales del código de alto nivel, Utilizar intrínsecos en el código del kernel. También desactivamos el software Prefetching debido al hecho de que los datos se encuentran en caches Después de calentarse.

6- Comparaciones 2013-2016: El co-procesador Xeon-Phi es un acelerador para un sistema host más tradicional, de la misma manera que Las GPU se utilizan conjuntamente con las CPU del host. Está disponible como una tarjeta complementaria estándar Pci-express para PCs. Las transferencias de datos del host al procesador y viceversa se producen a lo largo de 16 carriles de datos Pci-express (Gen. 2). La primera versión de producción de este sistema, que se espera esté disponible para el mercado de masas a finales de enero 2013, tiene un procesador MIC de Knights Corner (KNC) y 8 GBytes de memoria RAM GDDR5. El procesador KNC integra Hasta 61 núcleos de CPU, y se ejecuta a una frecuencia de ≈1 GHz. Se conecta a su banco de memoria privada a través de 16 Canales de memoria, proporcionando un ancho de banda máximo teórico de ≈320 GByte / s. El ancho de banda máximo teórico para El procesador host es ≈ 8 GByte / s.[2]

Cada núcleo se basa en la arquitectura Pentium, e incluye 32 KB de caché L1 utilizado para datos e instrucciones, 512 KB de caché de datos L2 y una unidad de punto flotante (FPU) de 512 bits de ancho. El motor FPU realiza una Fused-multiplique-agrega la instrucción por ciclo del reloj, entregando un funcionamiento máximo de ≈32 GFlops en la sola precisión, Y ≈16 GFlops en doble precisión, si todos los elementos del vector de datos pueden utilizarse en todos los ciclos de reloj. En este caso, Todo el KNC es capaz de ofrecer un rendimiento máximo de alrededor de 2 y 1 Tflops, respectivamente, en simple y doble Precisión, respectivamente. Dentro del procesador, los núcleos están conectados a través de un anillo bidireccional de alta velocidad, Y los elementos de datos dentro de todos los cachés L2 son compartidos por todos los núcleos. El KNC ejecuta una versión ligera modificada de la Sistema operativo Linux; Cada núcleo admite la ejecución de hasta 4 subprocesos de Linux.

La versión de host de vadd se ejecutará si el sistema en tiempo de ejecución no puede acceder a la tarjeta MIC. El descargado Puede eventualmente desovar varios hilos para ejecutarse en todos los núcleos disponibles. El lenguaje de descarga disponible en el El entorno de software MIC está integrado con varios lenguajes de programación existentes (C ++, Fortran, Cilk, TBB, . . .), Y los modelos (openMP y OpenCL [1, 2]).

Núcleo paralelo: el código que se ejecuta en el procesador MIC debe permitir que todos los núcleos para trabajar en paralelo la explotación MIMD o SPMD multi-tarea paralelismo; En el primer caso la aplicación se descompone en varias subtareas Y cada uno es ejecutado por un núcleo diferente; En este último caso, que también se puede combinar con el anterior Uno, el conjunto de datos es típicamente particionado entre los núcleos, y cada núcleo ejecuta la misma tarea en diferentes Partes del conjunto de datos.

III. CONCLUSIONES

El enfoque de diseño multi-núcleo se está convirtiendo rápidamente en la forma preferida de mejorar aún más las prestaciones del procesador A pesar de que las actuales tecnologías microelectrónicas ponen un límite superior práctico a la frecuencia de reloj A aproximadamente 3 GHz. Un procesador multi-núcleo es un único chip que integra dos o más CPUs independientes. Los Número de núcleos dentro de un chip está creciendo rápidamente: los procesadores con 100 o más núcleos se esperan en el futuro. El enfoque de muchos núcleos permite a los procesadores escalar según la ley de Moore, pero tiene un gran impacto En el diseño de las aplicaciones, lo que hace que el desafío de mantener el rendimiento de hardware a algoritmos y software

REFERENCIAS

[1]. Jianbin Fang, Ana Lucia Varbanescu, Henk Sips, Lilun Zhang, Yonggang Che, Chuanfu Xu "An empirical study of Intel Xeon Phi" Dec. 2013.

[2]. http://www.sciencedirect.com/science/article/pii/S1877705091300362*1*