ANÁLISIS DE RENDIMIENTO DE LAS TRANSFERENCIAS DE MEMORIA Y SUBRUTINAS GEMM EN CLÚSTER DE GPU NVIDIA TESLA

Los clúster comerciales aumentados con aceleradores de aplicaciones están evolucionando hacia sistemas de cómputo de alto rendimiento. Las unidades de procesamiento gráfico (GPU) con un alto costo son una buena plataforma para la aceleración de la aplicación científica.

INTRODUCCIÓN

- Entre las ***GPUs CUDA*** de ***NVIDIA*** disponibles, la serie Tesla esta diseñada específicamente para el campo de la computación científica. En este trabajo, se estudió el rendimiento de las copias de memoria y subrutinas GEMM, Para ello, un punto de referencia ***FRAMEWORK NETPIPE*** [1], que se ha desarrollado para evaluar la latencia y ancho de banda en las copias de memoria entre el ***HOST*** y el dispositivo ***GPU.***
- Un problema importante que aparece en el modelo de coprocesador es la sobrecarga involucrada en la transferencia de datos al espacio de memoria del dispositivo contra el tiempo de calculo actual.
- Uno de los principales objetivos del entorno de programación ***CUDA*** es el desarrollo de programas paralelos escalables y eficientes [4]. En este modelo, la ***GPU*** es vista como un dispositivo de computo multiprocesos que es capaz de ejecutar hilos en paralelo.

PALABRAS CLAVES

- **GPU NVIDIA TESLA** [5]
- **CPU** [6]
- **CLÚSTER** [7]
- **NETPIPE** [8]
- **SUBRUTINA** [9]
- **HOST** [10]
- **CUDA** [11]
- **HPC** [12]
- **GPU** [14]

GPU

Es un coprocesador. Se trata de un componente muy parecido al CPU, solo que el tipo de procesamiento al que se dedica es al de gráficos. [14]

http://nvidia.rave.com/product/nvidia-tesla-k40-gpu-accelerator/

CPU

es el hardware dentro de una computadora u otros dispositivos programables, que interpreta las instrucciones de un programa informático mediante la realización de las operaciones básicas aritméticas. [6]

http://definicion.mx/cpu/

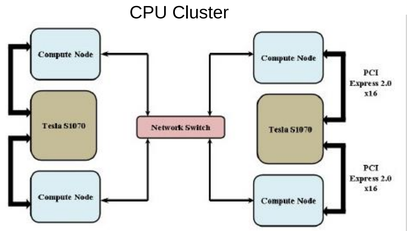
¿QUE ES?

Las GPU NVIDIA Tesla, son tarjetas de “video” como es común escuchar en el mercado, pero esta serie está específicamente diseñada para el campo de la computación científica.

En este artículo queremos mostrar que es una GPU NVIDIA Tesla, por qué se diferencia de sus contrapartes y como es el rendimiento de un ***cluster*** equipado con GPU NVIDIA TESLA, con respecto al rendimiento de las copias de memoria y subrutinas GEMM que son cruciales para aplicación de algoritmos de química computacional, lo que se logrará usando herramientas para analizar y evaluar la latencia y ancho de banda de las copias de memoria entre el ***Host*** y el dispositivo GPU.

Para medir el rendimiento que tienen el host-CPU y el dispositivo GPU, se basan en un ***framework*** llamado NetPIPE para evaluar la latencia y ancho de banda de las copias de la memoria entre el host y el dispositivo de la GPU.

Fig.2. ***Cluster of compute nodes connected to the Tesla S1070***



A continuación se muestra el clúster de GPU, éste está configurado con nodos de computación que comprenden procesadores Intel Xeon y servidores blade Tesla S1070 como se muestra a continuación en la Figura 2.

Se compara el funcionamiento de las subrutinas GEMM de precisión ***single*** y ***double*** de la biblioteca de NVIDIA CUBLAS 2.0 y se comparan con los resultados de las rutinas BLAS desde el procesador Intel Math Kernel Library (MKL), para entender ventajas y desventajas.

RESULTADOS

Evaluación de rendimiento y latencia de la memoria.

- La copias de memoria del ***host*** al dispositivo tienen un incremento lineal en el rendimiento que alcanzan un máximo de 6MB en el tamaño del búfer. No se evidencia una variación que pueda generar perturbación.
- Las transferencias de memoria del dispositivo al ***host*** alcanza un valor máximo de 2.0GB/s alrededor de 64MB. El comportamiento en el rendimiento es similar a la copia del host hacia el dispositivo.
- Para las copias de memoria ***ping-pong*** entre la memoria paginada del ***host*** y el dispositivo, se observa un rendimiento máximo de 1.6GB/s. Para un búfer incrustado en el host el rendimiento es de 2.9GB/s.

Rendimiento del SGEMM y DGEMM en una CPU

- Con operaciones GEMM, la escala de rendimiento es lineal con respecto al número de hilos de tal manera que las subrutinas son bien paralelizadas teniendo una buena utilización de los núcleos del procesador.

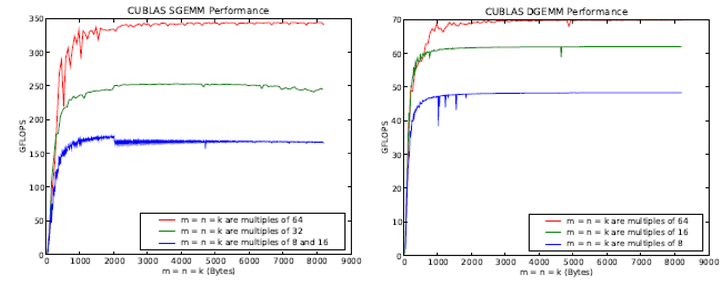


Fig. 8. (a) cublasSgemm performance for different matrix sizes (b) cublasDgemm performance for different matrix sizes

CONCLUSIONES

- Se estudió el rendimiento de las subrutinas SGEMM / DGEMM de la biblioteca CUBLAS 2.0 en la última versión del sistema de cómputo NVIDIA Tesla. Durante el desarrollo del módulo ***NetPIPE cudaMemcpy*** los resultados fueron validados con referencia a los resultados del ***benchmark*** dados por el ***CUDA SDK.***
- Basado en los resultados del estudio se llega como conclusión que el módulo ***cudaMemcpy*** de ***NetPIPE*** se puede utilizar para medir los datos de rendimiento de movimiento entre el host y el dispositivo ***GPU.***
- Como trabajo futuro, se tiene como idea principal integrar este módulo con el módulo ***NetPIPE Infiniband*** para incorporar copias de memoria a dispositivos remotos utilizando el GPU directo a memoria remota, se pretende también ampliar el ***framework NetPIPE*** para incorporar el rendimiento de las copias de memoria asíncronos.

Referencias

- [1] Q. O. Snell, A. R. Mikler, and J. L. Gustafson, “Netpipe: A network protocol independent performace evaluator,” in In Proceedings of the IASTED International Conference on Intelligent Information Manage-ment and Systems, 1996.
- [4] J. Nickolls, I. Buck, M. Garland, and K. Skadron, “Scalable parallel programming with cuda,” Queue, vol. 6, no. 2, pp. 40–53, 2008.
- [5] http://www.nvidia.es/object/tesla-server-gpus-es.html
- [6] https://es.wikipedia.org/wiki/Unidad_central_de_procesamiento
- [7] https://es.wikipedia.org/wiki/Cl%C3%BAster_(inform%C3%A1tica)
- [8] Framework para HPC – Performance Comparasion
- [9] https://es.wikipedia.org/wiki/Subrutina
- [10] https://es.wikipedia.org/wiki/Host
- [11] http://www.nvidia.es/object/cuda-parallel-computing-es.html
- [12] https://es.wikipedia.org/wiki/Computaci%C3%B3n_de_alto_rendimiento
- [13] https://es.wikipedia.org/wiki/Unidad_de_procesamiento_gr%C3%A1fico
- [14] http://www.nvidia.es/object/gpu-computing-es.html