¿Qué ventajas tiene un servidor dedicado con una o varias GPU?

¿Qué ventajas tiene un servidor dedicado con una o varias GPU? ¿Cuales son sus beneficios? ¿A quienes les sirve?
Con el auge de la Inteligencia artificial, el machine learning, la explosión del Big Data y el Data Science, uno se empieza a preocupar por la capacidad de cómputo de los servidores tradicionales, de sus alcances y posibles limitaciones.

Es una realidad que el aprendizaje automático y el Deep learning, no son una moda, son una tendencia.

¿Cómo podemos afrontar este tipo de retos coyunturales?

Podemos incorporar a nuestros servidores habituales procesadores más potentes pero siempre nos vamos a estrellar contra una barrera técnica: sólo hay un puñado de ALU (Unidades lógicas algorítmicas) en una CPU y la mayoría de transistores se emplean para el almacenamiento de datos de caché y el control de flujo. ¿Cómo superar esta limitante? ¿Llamamos a la liga de la justicia o los superamigos?
No es necesario llegar a tanto. Para eso están las GPU (es un acrónimo de Graphics Processing Unit). Un momento, ¿estamos hablando de las tarjetas de video ? Por supuesto. Pero vamos por partes.

El comienzo

Todo empezó cuando Nvidia, el principal proveedor mundial de GPU, se dio cuenta que su producto podía hacer mucho más que

renderizar videos e imágenes, que podía ayudar a resolver problemas matemáticos de computación intensiva.
Además, podrían hacerlo de forma más rentable que la supercomputación tradicional en las CPU. Debido a esto, desarrollaron la GPU de propósito general y un lenguaje de programación llamado CUDA para que los científicos escriban un código que interactúe directamente con la GPU. Ahora, algunas de las súper computadoras más poderosas del mundo, como Titan en el Laboratorio Nacional de Oak Ridge y Tianhe-1 en el Centro Nacional de Supercomputación en Tianjin, China, se construyen utilizando GPU.

Mas recientemente, la comunidad de aprendizaje automático ha comenzado a adoptar sistemas basados en GPU para algunos de sus problemas más desafiantes. La capacitación de estos algoritmos puede llevar tradicionalmente semanas o incluso meses en sistemas basados en CPU. Sin embargo, en los sistemas basados en GPU, ahora pueden entrenar en días o incluso horas. Los grandes grupos de investigación en Google, Baidu, Yahoo, Microsoft y Facebook están realizando un trabajo increíble en los campos de la visión artificial y el reconocimiento de voz al aprovechar las GPU y el aprendizaje profundo.

Donde la GPU supera a la CPU

Resulta que en una GPU, una gran parte de los transistores son ALU, dedicados al procesamiento de datos. De hecho, una nVIDIA Tesla K80 tiene 4992 núcleos CUDA para su procesamiento. Estas ALU son núcleos de subprocesos múltiples, simples, de datos paralelos, que ofrecen alta potencia de cómputo y un gran ancho de banda de memoria, todo con un consumo de energía muy bajo.

Donde no lo logra

Por diseño, una GPU podrá procesar datos varias veces más rápido que cualquier CPU, sin embargo, existen algunas limitaciones. Debido al diseño, el procesamiento secuencial en serie es menos efectivo en una GPU que en una CPU. Además, el desarrollo de algoritmos para GPU es complicado y requiere una programación sofisticada de bajo nivel. Hay algunos algoritmos que simplemente no pueden ser paralelizados debido a las interdependencias de datos inherentes en el algoritmo. Por lo tanto, un sistema heterogéneo con CPU y GPU puede proporcionar lo mejor de ambos mundos, procesamiento secuencial en serie y procesamiento altamente paralelizado.
Imagine un sistema heterogéneo de GPU y CPU que transmite datos casi en tiempo real a través de un sistema basado en GPU que puede realizar cálculos y procesamiento sobre la marcha para cualquier decisión rápida. Los datos se pueden guardar en la memoria por un corto período de tiempo para que los analistas puedan interactuar con los datos extremadamente rápido y realizar cálculos y transformaciones en un instante.

Cuatro criterios para el éxito con la tecnología GPU

1. La capacidad de ejecutar consultas en conjuntos de datos masivos: escala fácilmente de 500 gb a 40 tb
2. Alto rendimiento: reduce el tiempo de consulta de horas o días a segundos y minutos
3. Huella de hardware a pequeña escala y bajo costo: dos máquinas pueden hacer el trabajo de ocho bastidores
4. Fácil de usar: no se necesita capacitación especializada para tener éxito.

Conclusiones

Dicho lo anterior podemos hilvanar fino y afirmar que un servidor dedicado que implemente 1-10 GPU a su núcleo de procesamiento, estará en condiciones de manejar volúmenes de información y cálculos de mejor forma que un server ordinario; así mismo le añade potencial, por supuesto para renderizado de imágenes y video, para meterse de lleno en la predicción de datos, analogías y un largo etcétera.
Por citar algunas de las bondades que traen consigo: procesamiento 3-D rápido, aritmética de punto flotante precisa y procesamiento de números sin errores. Aunque normalmente las GPU funcionan a velocidades de reloj más lentas, tienen miles de núcleos que les permiten ejecutar miles de subprocesos individuales simultáneamente, como procesos computacionales significativos.
La ejecución de tareas intensivas en computación en una CPU puede comprometer todo el sistema. Descargar parte de este trabajo a una GPU es una excelente manera de liberar recursos y mantener un rendimiento constante.

Curiosamente, puede enviar las cargas de trabajo más difíciles a su GPU mientras la CPU maneja los principales procesos secuenciales. Tales estrategias GPGPU son críticas para brindar mejores servicios que atiendan a los usuarios finales, que experimentan un rendimiento acelerado.

Big Data prospera en entornos paralelos

Muchas de las tareas de Big Data que crean valor comercial implican realizar las mismas operaciones repetitivamente. La gran cantidad de núcleos disponibles en el servidor de servidores de GPU le permite realizar este tipo de trabajo dividiéndolos entre procesadores para procesar datos voluminosos a un ritmo más rápido.

Consumo de energía mejorado

No tiene que ser una empresa con conciencia ecológica para beneficiarse de la computación que ahorra energía. Los sistemas equipados con GPU que usan menos energía para realizar las mismas tareas imponen demandas menores a los suministros que los alimentan. En casos de uso específicos, una GPU puede proporcionar la misma capacidad de procesamiento de datos de 400 servidores solo con CPU.

Compatibilidad de software

Muchos paquetes de software modernos soportan la aceleración GPGPU. Algunos incluso le permiten paralelizar su código existente al incluir sugerencias que le indican al compilador dónde descargar el trabajo a la GPU. Por supuesto, es posible que necesite optimizar ciertas partes de sus aplicaciones, pero cuando es tan fácil aprovechar las ventajas de la computación en paralelo, no hay razón para detenerse.

Usted puede dar a sus procesos de aprendizaje de máquina un Head Start

Las tareas que se basan en el aprendizaje profundo y otros métodos de entrenamiento de IA se benefician enormemente de esta simbiosis. Los dispositivos de servidor dedicados de GPU pueden alimentar algoritmos de desarrollo de grandes volúmenes de datos en paralelo. Esta capacidad hace que sea mucho más fácil enseñarle a su software cómo reconocer las tendencias y los patrones que le interesa analizar.

Leer también: Desventajas y contras del Big Data ; Beneficios principales del Big Data