
oneAPI

¿Qué es oneAPI?

A medida que las cargas de trabajo de procesamiento intensivo se vuelven cada vez más diversas, también deben serlo las arquitecturas en las que se ejecutan. Desde los clústeres HPC hasta la IA y el aprendizaje automático, lograr el mayor rendimiento de las aplicaciones centradas en datos de hoy en día, requiere el desarrollo y la implementación en una combinación de motores de computación: CPU, GPU, FPGA y aceleradores especializados.

Para afrontar esto, Intel te ofrece oneAPI, un modelo de programación unificado y basado en estándares para simplificar el desarrollo entre arquitecturas y mejorar la eficiencia y la innovación.

OneAPI es entonces un modelo de programación abierto y basado en estándares que les permite a los desarrolladores usar una sola base de código en varias arquitecturas: CPU, GPU, FPGA y otros aceleradores. El resultado es una informática más rápida sin tener que depender de proveedores.

OneAPI básicamente consta de DPC++ y un conjunto básico de librerías. Los programas se escriben en C++, implementan el modelo de programación paralela SYCL y se compilan con un compilador DPC++.

SYCL (Standard C++ for Heterogeneous Parallel Programming) es una capa de abstracción multiplataforma que permite a los algoritmos cambiar entre aceleradores de hardware, como CPUs, GPUs y FPGAs, sin cambiar ni una sola línea de código. SYCL es un estándar abierto sin regalías desarrollado por el Khronos Group que permite a los desarrolladores programar arquitecturas heterogéneas en C++ estándar. Además, su modelo de programación utiliza un único origen, permitiendo que tanto el código principal como el del kernel se escriban en un solo archivo fuente.

DPC++ (Data Parallel C++) es una extensión del lenguaje C++ desarrollada por Intel para programación paralela y heterogénea. Esta extensión se basa en estándares abiertos como SYCL y está diseñada para aprovechar las capacidades de cómputo paralelo de las arquitecturas heterogéneas, como las GPUs y las CPUs.

Otros elementos de oneAPI

oneDPL: La biblioteca de oneAPI DPC++ proporciona la funcionalidad especificada en el estándar C++, con extensiones para admitir el paralelismo de datos y la transferencia a dispositivos, y con extensiones para simplificar su uso en la implementación de algoritmos de paralelismo de datos.

oneDNN: La Biblioteca de oneAPI de Redes Neuronales Profundas es una biblioteca de rendimiento que contiene bloques de construcción para aplicaciones y marcos de trabajo de aprendizaje profundo.

oneCCL: La Biblioteca de oneAPI de Comunicaciones Colectivas proporciona primitivas para los patrones de comunicación que ocurren en aplicaciones de aprendizaje profundo. oneCCL admite tanto la ampliación vertical para plataformas con múltiples dispositivos oneAPI como la ampliación horizontal para clústeres con múltiples nodos de cómputo.

Level Zero: Proporciona interfaces de bajo nivel directas al hardware que están adaptadas a los dispositivos en una plataforma oneAPI.

oneDAL: La Biblioteca de oneAPI de Análisis de Datos es una biblioteca que acelera el análisis de grandes conjuntos de datos al proporcionar bloques algorítmicos altamente optimizados para todas las etapas del análisis de datos (preprocesamiento, transformación, análisis, modelado, validación y toma de decisiones) en modos de procesamiento por lotes, en línea y distribuido.

oneTBB: la Biblioteca de oneAPI de Construcción de Hilos es un modelo de programación para la programación paralela escalable utilizando código estándar ISO C++. Un programa utiliza oneTBB para especificar el paralelismo lógico en algoritmos, mientras que una implementación de oneTBB asigna ese paralelismo a hilos de ejecución.

oneVPL: La Biblioteca de oneAPI de Procesamiento de Video es una interfaz de programación para la decodificación, codificación y procesamiento de video, con el fin de construir canalizaciones de medios portátiles en CPUs, GPUs y otros aceleradores. Proporciona descubrimiento y selección de dispositivos en cargas de trabajo centradas en medios y análisis de video, así como primitivas de API para compartir búferes sin copia.

oneMKL: La Biblioteca de oneAPI Kernel de Matemáticas define un conjunto de rutinas matemáticas fundamentales para su uso en cómputo de alto rendimiento y otras aplicaciones. Como parte de oneAPI, oneMKL está diseñada para permitir la ejecución en una amplia variedad de dispositivos computacionales: CPUs, GPUs, FPGAs y otros aceleradores. La funcionalidad se subdivide en varios dominios: álgebra lineal densa, álgebra lineal dispersa, transformadas discretas de Fourier, generadores de números aleatorios y operaciones matemáticas vectoriales.

Ray Tracing: Es un conjunto de rutinas avanzadas de trazado de rayos y renderización de alta fidelidad, así como cálculos, destinados a su uso en diversos ámbitos de gráficos 3D, que incluyen efectos visuales fotorrealistas para cine y televisión, renderización de animaciones, visualización científica, cálculos de alto rendimiento, videojuegos, etc.

Objetivo

“Uno de los principales problemas que enfrentan los desarrolladores hoy en día es la existencia de entornos de programación dispares y pocas oportunidades de reutilización de código en diferentes tipos de hardware. Un entorno de programación único que pueda ejecutar código sin sacrificar el rendimiento en diversos tipos de hardware es un desafío difícil pero importante. Intel oneAPI parece ser un paso significativo en la dirección correcta, prometiendo portabilidad de código sin comprometer la capacidad de ajustar el rendimiento para CPUs y aceleradores, y haciendo que las transiciones de hardware sean considerablemente menos arriesgadas y propensas a errores.” Federico Carminati, Director de Innovación

El principal objetivo de oneAPI radica en simplificar la reutilización del mismo código en diversas plataformas, incluso al emplear compiladores cruzados diferentes, todo ello mientras garantiza el máximo rendimiento necesario para la aplicación. Esto implica que, independientemente de los dispositivos y aceleradores instalados en el sistema, así como de los lenguajes y bibliotecas empleados por cada uno de estos elementos, ya sean middleware, frameworks, aplicaciones o cargas de trabajo, oneAPI trabaja para cohesionar, abstraer y brindar soporte a todos estos dispositivos en un terreno común.

En este contexto, oneAPI permite compartir espacios de memoria, portar y reutilizar código de manera eficiente, y posibilita que las herramientas operen sin problemas en diversas arquitecturas. Esta integración y abstracción facilitan la creación de aplicaciones versátiles y eficientes que pueden adaptarse a diferentes entornos de hardware y software, promoviendo así la flexibilidad y la eficacia en el desarrollo de soluciones computacionales.

Por lo que público esperado incluye:

- Desarrolladores de aplicaciones, desarrolladores de middleware, proveedores de software de sistema y proveedores de hardware.
- Desarrolladores y arquitectos de software C, C++, Data Parallel C++, Fortran, Phyton, OpenMP y MPI que crean soluciones HPC, Enterprise, AI y Cloud.
- Desarrolladores que buscan maximizar el rendimiento y flexibilidad de sus softwares para que sean compatibles con arquitectura cruzada en las plataformas actuales o futuras.

Arquitecturas

OneAPI proporciona una interfaz de desarrollo común en una variedad de aceleradores paralelos de datos.

Los programadores usan SYCL tanto para la programación API como para la programación directa. Las capacidades de una plataforma en OneAPI están determinadas por la interfaz Level Zero, que proporciona al software del sistema.

Intel tiene una posición única para ofrecer una mezcla diversa de arquitecturas escalares, vectoriales, de matriz y espaciales implementadas en sockets de CPU, GPU, aceleradores y FPGA. Esto brinda la capacidad de utilizar el tipo de cómputo más adecuado donde se necesita.

Intel utiliza 4 arquitecturas principales:

La arquitectura escalar se refiere típicamente al tipo de cargas de trabajo que son óptimas en una CPU, donde una secuencia de instrucciones opera a una velocidad determinada, generalmente impulsada por los ciclos del reloj de la CPU. Desde el arranque del sistema y las aplicaciones de productividad hasta cargas de trabajo avanzadas como la criptografía y la inteligencia artificial, las CPUs basadas en escalares trabajan en una amplia gama de topografías con un rendimiento constante y predecible.

La arquitectura vectorial es óptima para cargas de trabajo que pueden descomponerse en vectores de instrucciones o vectores de elementos de datos. Las GPU y las VPU ofrecen procesamiento paralelo basado en vectores para acelerar la representación gráfica en juegos, medios ricos, análisis y entrenamiento e inferencia de aprendizaje profundo. Al escalar arquitecturas vectoriales desde clientes, centros de datos y el borde, podemos llevar el rendimiento del procesamiento paralelo desde gigaFLOPS hasta teraFLOPS, petaFLOPS y exaFLOPS.

La arquitectura de matriz obtiene su nombre de una operación común realizada típicamente para cargas de trabajo de inteligencia artificial (multiplicación de matrices). Mientras que otras arquitecturas pueden ejecutar código de multiplicación de matrices, los ASIC han logrado tradicionalmente el rendimiento más alto implementando el tipo de operaciones típicamente necesarias para la inferencia y el entrenamiento de inteligencia artificial, incluida la multiplicación de matrices.

La arquitectura espacial es una arquitectura especial generalmente asociada con una FPGA. Aquí, los datos fluyen a través del chip, y la operación de cómputo realizada en el elemento de datos se basa en la ubicación física de los datos en el dispositivo. El algoritmo de transformación de datos específico que se ha programado en la FPGA.

Intel, esta planificando las arquitecturas del futuro con investigación y desarrollo en cómputo de próxima generación. Entre estas se encuentran las **arquitecturas cuánticas y neuromórficas**.

Aplicaciones

Deep Learning y Redes Neuronales: El marco de trabajo oneAPI permite a los desarrolladores implementar y optimizar modelos de deep learning en hardware diverso, incluyendo CPUs, GPUs y otros aceleradores.

Procesamiento de Señales e Imágenes: oneAPI es utilizado en aplicaciones de procesamiento de señales y de imágenes, como el procesamiento de imágenes médicas, la mejora de imágenes y el análisis de señales, aprovechando las capacidades de aceleración de hardware.

Simulación y Modelado: En campos como la ingeniería y la simulación, oneAPI ayuda a desarrollar y optimizar modelos complejos para la simulación de fenómenos físicos, como dinámica de fluidos, simulación de materiales y diseño de productos.

Ciencia de Datos: En aplicaciones de ciencia de datos y análisis, oneAPI facilita la implementación eficiente de algoritmos en hardware heterogéneo para procesar grandes conjuntos de datos y realizar análisis avanzados.

Juegos y Entretenimiento: En la industria de los videojuegos, oneAPI puede utilizarse para desarrollar y optimizar gráficos, físicas y otras tareas intensivas en cómputo, mejorando la experiencia del usuario.

Internet de las Cosas (IoT): oneAPI es aplicable en el desarrollo de soluciones para el Internet de las Cosas, donde la eficiencia energética y el rendimiento son críticos, especialmente cuando se trabaja con dispositivos IoT que tienen hardware heterogéneo.

Conclusión

El futuro de oneAPI se presenta como una evolución significativa en el panorama de la programación heterogénea. Esta iniciativa liderada por Intel ha demostrado su capacidad para unificar la programación en una amplia gama de dispositivos, desde CPUs hasta GPUs y FPGAs, ofreciendo a los desarrolladores una plataforma versátil y eficiente. A medida que la computación heterogénea se convierte en la norma en la era de la informática avanzada, oneAPI emerge como una solución integral que permite a los programadores aprovechar al máximo el rendimiento diversificado de los dispositivos de hardware modernos.

La promesa de escribir código una vez y ejecutarlo en una variedad de arquitecturas, sin comprometer el rendimiento, señala el potencial de oneAPI para simplificar el desarrollo de software en entornos heterogéneos. Con un enfoque centrado en estándares abiertos, oneAPI está posicionado para fomentar la colaboración industrial y la innovación, permitiendo que aplicaciones cruciales, desde la simulación científica hasta el mundo del entretenimiento, se beneficien de la eficiencia y la escalabilidad que ofrece.

A medida que la adopción de tecnologías heterogéneas continúa en aumento, se espera que oneAPI desempeñe un papel fundamental en la creación de soluciones de software eficientes y portátiles. Su capacidad para adaptarse a futuras innovaciones y tecnologías emergentes posiciona a oneAPI como una herramienta valiosa para los desarrolladores que buscan optimizar el rendimiento de sus aplicaciones en un mundo cada vez más diverso de hardware especializado.

Ejemplo

Utilizando un pequeño programa híbrido de OpenMP y MPI, a través de la plataforma con infraestructura de Cloud gratuita para los usuarios de Intel JupyterLab (https://jupyter.oneapi.devcloud.intel.com/), que facilita el acceso remoto a uno de los núcleos públicos de Intel a través de OneAPI, he desarrollado dos ejemplos simples. Los archivos CPP correspondientes y el cuaderno de JupyterLab pueden visualizarse en el siguiente enlace de GitHub:

https://github.com/Karim-Neme/Arquitectura-de-Computadoras-II---Final/

Bibliografía

https://www.intel.la/content/www/xl/es/developer/tools/oneapi/overview.html
https://es.wikipedia.org/wiki/OneAPI_(aceleraci%C3%B3n_de_c%C3%B3mputo)
https://www.oneapi.io/
https://www.youtube.com/@IntelSoftware
https://www.danysoft.com/intel-one-api/
https://spec.oneapi.io/versions/1.0-rev-2/index.html