CUDA es una arquitectura de cálculo paralelo de NVIDIA que aprovecha la gran potencia de la GPU para proporcionar un notable incremento del rendimiento del sistema. Esta tecnología nos permite realizar coprocesamiento repartido entre la CPU y la GPU, intentado explotar las ventajas de las GPU frente a las CPU de propósito general utilizando el paralelismo que ofrecen sus múltiples núcleos, permitiendo el lanzamiento de un alto número de hilos simultáneos.

 CUDA ante los otros modelos GPGPU es que su arquitectura, está diseñada para usar eficientemente de computaciones no-gráficas en la GPU y usa C. Soporta operaciones tanto binarias como enteras y no requiere convertir algoritmos en un formato pipeline.

Ofrece funcionalidades del hardware como la memoria compartida. Es una pequeña cantidad de memoria que puede ser accedida por bloques de threads. Permitiendo guardar los guardar los datos utilizados más frecuentemente.

Las ventajas que nos presenta CUDA son:

-El interfaz de programación de CUDA está basado en C con extensiones que facilitan el aprendizaje de CUDA.

-Podemos consultar cualquier posición de la memoria.

-Se dispone de un área de memoria que se compartida entre threads.

-Lecturas más rápidas de y hacia la GPU.

-Soporte hardware para operaciones binarias y enteras.

CUDA también presenta una serie de limitaciones:

-No se puede usar recursividad, punteros a funciones, variables estáticas en funciones.

-Puede existir un cuello de botella entre la CPU y la GPU por los anchos de banda de los buses y sus latencias.

-Los Hilos de ejecución (threads) deben lanzarse en grupos de al menos 32.

-Arquitectura cerrada, pertenece a NVIDIA.

Aplicaciones que se pueden beneficiar de estas tecnologías serian edición de video y fotografía, creaciones de imágenes tridimensionales, diseño de objetos, diseño de nuevos materiales y estudio de estos, entre otras muchas más aplicaciones posibles.