\chapter{GPU With CUDA Architecture Updates}

\label{sec:gpu\_architecture}

A GPU is a specialized computer that provides fast image processing and construction for display. However, over a decade ago NVidia started to adjust the architecture for general applications that could be executed in parallel. The NVidia GPU consists of multiple clusters of core arrays, with each core array sharing cache memory and registers for execution.

This architecture allows us to run programs in the mode ``same instruction, multiple data’’ (SMID) where, for each clock cycle, all active cores run the same operation but on different parts of the memory.