GPU- Acceleration

A graphics processing unit (GPU) is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device. GPUs are used in embedded systems, mobile phones, personal computers, workstations, and game consoles. Modern GPUs are very efficient at manipulating computer graphics and image processing. Their highly parallel structure makes them more efficient than general-purpose central processing units (CPUs) for algorithms that process large blocks of data in parallel. In a personal computer, a GPU can be present on a video card or embedded on the motherboard. In certain CPUs, they are embedded on the CPU die.

The term "GPU" was coined by Sony in reference to the PlayStation console's Toshiba-designed Sony GPU in 1994. The term was popularized by Nvidia in 1999, who marketed the GeForce 256 as "the world's first GPU". It was presented as a "single-chip processor with integrated transform, lighting, triangle setup/clipping, and rendering engines".[4] Rival ATI Technologies coined the term "visual processing unit" or VPU with the release of the Radeon 9700 in 2002.

GPU-accelerated computing is the employment of a graphics processing unit (GPU) along with a computer processing unit (CPU) in order to facilitate processing-intensive operations such as deep learning, analytics and engineering applications. Developed by NVIDIA in 2007, the GPU provides far superior application performance by removing processing-intensive application sections to GPU. GPU-accelerated computing deployment is growing in popularity due to the large variety of applications in which it could be used, such as artificial intelligence, drones, robots or autonomic cars.

The GPU helps in providing superior performance for software applications. From the perspective of the user, GPU-accelerated computing makes applications faster. GPU-accelerated computing functions by moving the compute-intensive sections of the applications to the GPU while remaining sections are allowed to execute in the CPU. While the CPU is comprised of cores designed for sequential serial processing, the GPU is designed with a parallel architecture consisting of more efficient yet smaller cores that can easily handle multiple tasks in parallel. As a result, in GPU-accelerated computing, while sequential calculations are performed in the CPU, highly complicated calculations are computed in parallel in the GPU. Another salient feature of GPU-accelerated computing is the support offered to all the parallel programming models, thus helping application designers and developers to provide superior application performance.

GPU-accelerated computing has been extensively used in video editing, medical imaging, fluid simulations, color grading and enterprise applications, and its use is promising in complex fields such as artificial intelligence and deep learning.

Modern GPUs use most of their transistors to do calculations related to 3D computer graphics. In addition to the 3D hardware, today's GPUs include basic 2D acceleration and framebuffer capabilities (usually with a VGA compatibility mode). Newer cards such as AMD/ATI HD5000-HD7000 even lack 2D acceleration; it has to be emulated by 3D hardware. GPUs were initially

used to accelerate the memory-intensive work of texture mapping and rendering polygons, later adding units to accelerate geometric calculations such as the rotation and translation of vertices into different coordinate systems. Recent developments in GPUs include support for programmable shaders which can manipulate vertices and textures with many of the same operations supported by CPUs, oversampling and interpolation techniques to reduce aliasing, and very high-precision color spaces. Because most of these computations involve matrix and vector operations, engineers and scientists have increasingly studied the use of GPUs for non-graphical calculations; they are especially suited to other embarrassingly parallel problems.

With the emergence of deep learning, the importance of GPUs has increased. In research done by Indigo, it was found that while training deep learning neural networks, GPUs can be 250 times faster than CPUs. The explosive growth of Deep Learning in recent years has been attributed to the emergence of general purpose GPUs. There has been some level of competition in this area with ASICs, most prominently the Tensor Processing Unit (TPU) made by Google. However, ASICs require changes to existing code and GPUs are still very popular.

Most GPUs made since 1995 support the YUV color space and hardware overlays, important for digital video playback, and many GPUs made since 2000 also support MPEG primitives such as motion compensation and iDCT. This process of hardware accelerated video decoding, where portions of the video decoding process and video post-processing are offloaded to the GPU hardware, is commonly referred to as "GPU accelerated video decoding", "GPU assisted video decoding", "GPU hardware accelerated video decoding" or "GPU hardware assisted video decoding".

More recent graphics cards even decode high-definition video on the card, offloading the central processing unit. The most common APIs for GPU accelerated video decoding are DxVA for Microsoft Windows operating system and VDPAU, VAAPI, XvMC, and XvBA for Linux-based and UNIX-like operating systems. All except XvMC are capable of decoding videos encoded with MPEG-1, MPEG-2, MPEG-4 ASP (MPEG-4 Part 2), MPEG-4 AVC (H.264 / DivX 6), VC-1, WMV3/WMV9, Xvid / OpenDivX (DivX 4), and DivX 5 codecs, while XvMC is only capable of decoding MPEG-1 and MPEG-2.

The video decoding processes that can be accelerated by today's modern GPU hardware are:

- Motion compensation (mocomp)
- Inverse discrete cosine transform (iDCT)
- Inverse telecine 3:2 and 2:2 pull-down correction
- Inverse modified discrete cosine transform (iMDCT)
- In-loop deblocking filter
- Intra-frame prediction
- Inverse quantization (IQ)
- Variable-length decoding (VLD), more commonly known as slice-level acceleration
- Spatial-temporal deinterlacing and automatic interlace/progressive source detection
- Bitstream processing (Context-adaptive variable-length coding/Context-adaptive binary arithmetic coding) and perfect pixel positioning.