Chapter 6
History and Evolution of GPU Architecture

Prashanta Kumar Das
Govt. ITI Dhansiri, Barpathar, India

Ganesh Chandra Deka
Regional Vocational Training Institute for Women, India

ABSTRACT

The Graphics Processing Unit (GPU) is a specialized and highly parallel microprocessor designed to offload 2D/3D image from the Central Processing Unit (CPU) to expedite image processing. The modern GPU is not only a powerful graphics engine, but also a parallel programmable processor with high precision and powerful features. It is forcasted that by 2020, 48 Core GPU will be available while by 2030 GPU with 3000 core is likely to be available.This chapter describes the chronology of evolution of GPU hardware architecture and the future ahead.

1. INTRODUCTION

Graphics on desktop computers were handled by Video Graphics Array (VGA) controller. A VGA controller is simply a memory controller attached to DRAM and a display generator. The main function of a VGA is to receive image data, arrange it properly, and send it to a video device, such as a computer monitor for display. By the end of 1990s, different graphics acceleration components were being added to the VGA controller for rasterizing triangles, texture mapping, and simple shading. NVIDIA(a corporation in Santa Clara, California, USA) released the “GeForce 256” and marketed as the world’s first GPU in the year of 1999.

Like Central Processing Unit (CPU), GPU is a single-chip processor. The difference between the CPU and GPU is that, GPU may have hundreds of “Core”s while the number of “Core”s in CPU is comparatively very less (4 or 8 Cores). The main purpose of the GPU is to compute 3D functions resulting to requirement of comparatively higher number of “Core”s . As 3D calculations are extremely heavy GPU can help the computer run not only faster but also more efficiently.

DOI: 10.4018/978-1-4666-8853-7.ch006

GPU came into existence for graphical purpose, it has now evolved into computing, accuracy and performance. The fast development of the GPU, over the last few years has opened up a new world of possibilities for high-speed computation, ranging from biomedical to computer vision applications. GPU is the future of computation. Graphics cards are widely used for accelerated rendering of 3D scenes and in the field of image processing. The computational capability of the GPUs are mostly used in parallel computing units, since it is simple to program a graphics processor to perform general parallel tasks. GPU computation is high-speed as compared to CPU computation; thus it is one of the most exciting areas of research in the field of modern industrial research and development. GPU permits to run high definition graphics on computers, which are the demand of modern computing.

A GPU is a dedicated parallel processor optimized for accelerating graphical computations. The GPU is designed specifically to do the many floating-point calculations essential to 3D graphics processing. The development of GPU hardware architecture was started with a specific single core, fixed function hardware, pipeline implementation made solely for graphics, to a collection of extremely parallel and programmable cores for general purpose computation. The development in GPU technology has kept adding more programmability and parallelism to GPU core architecture resulting to a more general purpose CPU-like core. Modern GPU are particularly parallel, and are fully programmable ("Video card," 2010).

GPUs can be set up in a broad range of organizations, from desktops and laptops to mobile handsets and super computers. With their parallel structure, GPUs implement a variety of 2D and 3D graphics primitives processing in hardware, making them a general purpose central processing unit for these operations ("Graphics processing unit," 2010).

The original GPUs were modeled once the idea of a graphics pipeline. The graphics pipeline could be an abstract model of stages that graphics data is sent through, and is typically enforced via a combination of hardware (GPU cores) and central processing unit (CPU) package (OpenGL, DirectX). The graphics pipeline design approach is fairly uniform among the most important GPU makers like NVIDIA, ATI, etc., and helped accelerate GPU technology adoption. The pipeline simply transforms the coordinates of 3D images (determined by the software engineer) into 2D pixel images on the screen. It is essentially an "assembly line" of various stages of operations to apply to pixels/triangles, and can be generalized out of 2 stages:

- Geometry and Rendering.

Early GPUs enforced solely the rendering stage in hardware, requiring the CPU to produce triangles for the GPU to work on. As GPU technology progressed, more and more stages of the pipeline were enforced in hardware on the GPU, releasing up additional CPU cycles.

2. GPU EVOLUTION

1980's

Going up to the first 1980's, the "GPUs" of the time were very simply integrated frame buffers. They were boards of TTL(Transistor–Transistor Logic (TTL), a class of digital circuits built from Bipolar Junction Transistors (BJT) and resistors. In TTL both the logic gate function e.g., AND and the amplifying