

GPU Computing

Weijie Zhao

09/12/2023

GPU (Graphics Processing Unit)

- Rendering
 - 3D surfaces
 - Textures
 - Lights
 - Views







VIA 9GAG.COM



GPU Rendering

- Direct3D
- OpenGL

- Use primitives to render a frame

GPU ~~Rendering~~ Computing (Before 2007)

- Direct3D
- OpenGL

- Use primitives to render a frame
- **Make your 2D array as a frame and call render primitives**

GPU Computing (After 2007)

- CUDA (~~Compute Unified Device Architecture~~)
- C/C++, Fortran
- GPGPU (General-Purpose computing on Graphics Processing Units)
- OpenCL

GPU Architecture

- 108 Streaming Multi-processor (SM)
- 40 GB High-Bandwidth Memory (HBM)
 - 1555 GB/sec
 - 6912 FP32 CUDA cores
- 432 Tensor Cores, TensorFloat-32(TF32) Dense Tensor (156 TFLOPs)
- 192KB * 108 L1 Cache
- 40960 KB L2 Cache

GPU Scheduling

- SIMT (Single Instruction Multiple Thread)
- Warp
- Dangerous to implement critical section (Pre Volta)

- Independent Thread Scheduling (After Volta)