Gauss-Seidel Smoother

- Solving PDE

- Parallel
- Synchronization
- Lock
- Communication

Misc

- -O2 -O3
- -march=native
- Measure speedup
- Find bottleneck
  - Computation
  - Memory
  - Synchronization
- gprof
  - -g –gp
- gperftools

GPU Computing

Weijie Zhao

01/23/2025

GPU (Graphics Processing Unit)

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

https://en.wikipedia.org/wiki/File:Stanford_bunny_qem.png

https://en.wikipedia.org/wiki/Utah_teapot

https://www.quora.com/What-has-a-better-story-GTA-V-or-GTA-San-Andreas

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)