Paper Review

GPU Computing

Yueyang Li

2024-01-30

1. Paper Title, Authors, and Affiliations

- **Title**: GPU Computing
- **Year**: 2008
- **Authors**: John D. Owens, Mike Houston, David Luebke, Simon Green, John E. Stone, James C. Phillips
- **Affiliations**: University of California Davis, Stanford University, NVIDIA Corporation, University of Illinois at Urbana-Champaign

2. Main Contribution of the Paper

This paper provides a comprehensive overview of GPU computing, detailing several key contributions:

- The evolution of GPUs from fixed-function graphics processors to programmable, massively parallel computing devices
- Analysis of the advantages of GPU computing in handling highly parallel workloads with high throughput
- Detailed exploration of GPGPU (General-Purpose computing on GPUs) and its impact on scientific computing and game physics
- Presentation of real-world GPU applications through case studies, including Havok FX for game physics and biophysical simulations

3. Outline of Major Topics

1. **Introduction**
   (a) GPUs have evolved into highly parallel computing units
   (b) Parallelism is key to performance improvements
   (c) GPU computing (GPGPU) as a viable alternative for high-performance computing
2. **GPU Architecture**

   (a) Graphics pipeline overview and parallel processing capabilities
   (b) Evolution of programmability from fixed-function to modern GPUs
   (c) Comparison of GPU vs. CPU architectures and priorities
3. **GPU Programming Models**
   (a) SPMD (Single Program Multiple Data) paradigm
   (b) Memory access patterns and efficiency considerations
   (c) Scatter/Gather operations capabilities
4. **Software Environments for GPGPU**
   (a) Early GPU programming challenges
   (b) Modern high-level programming models (CUDA, OpenCL, etc.)
   (c) Commercial and academic solutions
5. **GPU Algorithms and Applications**
   (a) Essential GPU operations (Map, Reduce, Scan, Sort)
   (b) Physics simulations and scientific computing applications
   (c) Case studies of Havok FX and biophysical simulations

4. Two Things Liked or Found Interesting

1. The paper provides an insightful historical perspective on GPU evolution, clearly explaining the transition from fixed-function graphics accelerators to programmable, massively parallel processors.
2. The real-world success stories, particularly Havok FX and Folding@Home, demonstrate impressive speedup factors (10× to 60× over CPUs) and highlight the practical impact of GPU parallelism.

5. What Did You Not Like About the Paper?

- The paper focuses heavily on compute power but does not deeply discuss memory bandwidth limitations, which can significantly affect GPU performance in certain applications.
- While the paper discusses the state of GPU computing as of 2008, it does not provide clear predictions for future evolution of GPUs, particularly missing the emergence of AI acceleration as a major GPU application.

6. Questions for the Authors

1. How can GPUs handle problems with significant inter-thread communication more effectively, particularly for tasks like graph algorithms and adaptive mesh refinement?
2. The original goal of GPU is not intended for General Parallel processing, as the importance of Generative AI is growing, will the current GPU design be specialized for those general purpose inference tasks? Will Seperate Units be designed that can provide local environment for tedious inference tasks?