Scalable Computing – Assignment 2

By: Bhaskar Rao, **Student Id**: 18300829, **Course Code:**CS7NS1, **Stream:** MSc CS Data Science, **Date:** 27 Sep 18

1. **Advances in Multi-GPU Smoothed Particle Hydrodynamics Simulations by** Eugenio Rustico, Giuseppe Bilotta, Alexis He´ rault, Ciro Del Negro, and Giovanni Gallo

Identify the five key contributions/findings/conclusions of the paper

- A load-balanced, scalable multi-GPU(CUDA tech) implementation of fluid-dynamic simulation using SPH model was presented
- The multi-GPU based simulator which has GPU load-balancing has higher overall efficiency than the one which doesn't
- Simulation of fluid particle flows has application across numerous fields ranging from aerodynamics to special effects. Computation fluid dynamics models are used based on problem specifications & accuracy requirements
- There is scope for research in improving load-balancing algorithms & development of multi-node simulator for GPU clusters

Identify the five key technology insights provided by the paper

- NVIDIA's CUDA architecture (2007) allows easier accessibility to GPU resources for general purpose computing
- Although GPUs have lower clock rates, they have a large number of cores (in 1000) in comparison to a CPU; the hardware architecture of a GPU makes it ideal for solving computationally intensive problems which demonstrate high parallelism
- When using multiple GPUs concurrently to execute computational fluid-dynamics(CFD) simulation, CDUA's asynchronous API was exploited for computations and data transfers between the GPUs to integrate their simulation
- NVIDIA GTX480 GPU launched in 2010 has 480 CUDA cores,1536 MB of Global memory & processor clock speed of 1401MHz

Identify the five key insights of relevance to CPU, GPU and processor scalability

- Scaling of multi-GPU systems is still in its adolescence & requires more research. load balancing requires more refinement.
- Utilizing multiple GPUs to run CFD simulations has two benefits; using multiple device allows for executing simulations having data larger than a single device's memory along with massive decrease in simulation completion time
- The optimal way to exploit multiple GPUs simultaneously in a simulation depends on the simulation method. For smoothed particle hydrodynamics (SPH) model, authors used spatial decomposition as the SPH method is parallel with spatial locality
- Authors' were able to demonstrate that simulations would scale linearly with the number of Graphical Processing Units used based on their multi-GPU SPH fluid simulator implementation

2. **Parallel GPU Architecture Simulation Framework Exploiting Architectural-Level Parallelism with Timing Error Prediction by** Sangpil Lee &Won Woo Ro

Identify the five key contributions/findings/conclusions of the paper

- A parallel GPU simulation scheme "error predictive synchronisation" was introduced in the paper which has an effective thread synchronisation. This simulation time is reduced making this scheme a helpful tool for architecture GPU research
- The present advanced GPUs used for high-performance computing have very complex architecture consisting of 1000s of processing elements. The current architecture simulators don't have the required speed to be used in architecture research
- The EPS scheme had a speed improvement of 8.9 times on 16 core GPU when compared with existing single-thread simulator
- The current simulators have long simulation turnaround times for GPUs having many cores. The slow speed proves to be a severe bottleneck during the GPU architecture design development, verification & performance analysis

Identify the five key technology insights provided by the paper

- Today's advanced GPUs are required to process vast amount of data. Thus, in order to improve their processing power, they integrate large number of compute units & execution units which leads to architectural complexity and further challenges
- AMD's "Southern Islands" family & NVIDIA's Kepler/Fermi/Tesla GPU series are standard modern GPU architecture
- Although high level execution model of NVIDIA & AMD is similar, their internal CU's hardware is very different
- By classifying the components of the GPU as shared & independent, parallel GPU simulation architectures can be built (synchronization required for shared components during simulation)

Identify the five key insights of relevance to CPU, GPU and processor scalability

- The conventional architecture simulators(sequential) have long simulation times when used for large-scale multi-core processors
- Slack simulation scheme allows parallel GPU core simulation, but performance scalability is limited only to 4 compute unit (CU) threads. For more than 8 CU threads, idle time increases due to thread synchronization leading to decreased performance
- Error predictive synchronization(EPS) scheme predicts the cycle-error & performs synchronisation accordingly. This selective thread synchronisation leads to better performance scalability (than slack scheme) for multi thread/multi core simulations
- The modern computing is moving towards multi-core architecture, the number of processing elements on a machine is scaling rapidly (powerful GPUs ~ 2000 elements). Parallel & high-speed architecture simulators are required to match this pace