**Title:** Architectural Approaches to Design Reliable and Energy-Efficient GPUs

**Student Name:** Jingweijia Tan

**Degree:** PhD in Electrical Engineering

**Location:** ECE Large conference room (D N328), Engineering Building I,

**Time:** 11:30am, March 7th, 2016

**Committee Chair:** Dr. Xin Fu, ECE

**Committee Members:**
Dr. Jinghong Chen, ECE
Dr. Yuhua Chen, ECE
Dr. Jiming Peng, IE
Dr. Guoning Chen, CS
Dr. Shuaiwen Leon Song, PNNL

**Abstract:**

Modern graphic processing units (GPUs) support thousands of concurrent threads and provide high computational throughput, which makes them popular platforms for general-purpose high-performance computing (HPC) applications. However this raises reliability and energy-efficiency challenges in GPU architecture design. Originally designed for graphics applications with relaxed requirements on execution correctness, GPUs lack the error detection and fault tolerance features. In contrast, HPC programs have rigorous demands on execution correctness, which poses serious reliability challenges for general purpose computing on GPUs (GPGPUs). In addition, GPUs consumes large amount of energy to achieve its high computing power. The peak power consumption of a high-end GPU is more than twice of the CPU counterparts and the energy-efficiency of GPUs fail to grow as fast as the performance improvement.

In this thesis, I introduce several architectural approaches to design reliable and energy-efficient GPUs. I first propose several opportunistic techniques to recycle the idle time of streaming processors for soft-error detection and obtain the good fault coverage with negligible performance degradation. Utilizing the promising benefits of resistive memory, I further propose to leverage resistive memory to enhance the soft-error robustness and reduce the power consumption of registers in the GPUs. I then explore to mitigate the susceptibility of GPU register file to process variations. The proposed techniques are able to significantly optimize GPUs' performance under process variations. After that, I propose an effective and low-cost mechanism to maintain the register file reliability with negligible performance loss under process variations and low supply voltages, which enables substantial energy savings via aggressive supply voltage reduction. Finally, I propose an energy-efficient GPU L2 cache design that leverages locality similarity to reduce the L2 energy consumption with negligible performance

degradation. Overall, these techniques efficiently address the reliability and energy-efficient challenges in GPU architectures.