*Electrical & Computer Engineering*

# D E F E N S E

**Louisiana State University**

## Cache-Conscious Sparse Matrix Dense Matrix Multiplication on GPUs

***a dissertation to be defended by***


## *Haoqiang Guo*

### Ph.D. Candidate

### LSU Division of Electrical & Computer Engineering



**Abstract**—Sparse matrix-dense matrix multiplication (SpMM) is central to many deep learning workloads, especially graph neural networks (GNNs). Although GPUs offer massive parallelism utilizing them for SpMM remains challenging due to the irregular structure of the sparse matrix used to represent graphs. Without care, dense matrix re-use potential will go unexploited, slowing execution due to avoidable data movement. Load imbalance is another pitfall. We introduce a cache-conscious GPU implementation of SpMM for GNN applications. It includes a scheduling framework that improves locality and balance, especially on re-numbered graphs. Our approach combines non-uniform tiling, which groups dense regions of nonzeros to increase L1 cache reuse, with a coordination strategy that enables data sharing across thread blocks. Evaluated on GNN-derived matrices on an NVIDIA H100 GPU, it significantly outperforms optimized cuSPARSE baselines.


**When:** **Monday, 15 December 2025, 12:30 - 13:30 (Public Portion)**
**Where:** `https://lsu.zoom.us/j/93727247227`
**Info:** `https://www.lsu.edu/eng/ece/seminar`