# Chengming Zhang

+1 2052390936 | chengming.zhang@wsu.edu | Pullman, WA

## EDUCATION

**Washington State University** 08.2020 - Present
PhD, Computer Science GPA 3.85/4.0
**The University of Alabama** (transfer to WSU) 08.2019 - 08.2020
PhD, Computer Science GPA 4.0/4.0
**University of Electronic Science and Technology of China (UESTC)** 09.2013 - 06.2017
Bachelor of Engineering, Integrated circuit design and integrated system GPA 3.81/4.0

## WORKING EXPERIENCE

**DeepSpeed team at Microsoft** 05.2022 - 08.2022
Research Internship
- Optimize the training speed of large-scale recommended system.

**Facebook Reality Labs** 06.2021 - 08.2021
Engineer Internship
- Design scripts to generate efficient C++ codes according to users' requirements.
- Implement user-friendly UI to show statistics for the data read from the database.
- Analyze and optimize the original system call flow to implement a fully automated call flow.

## ACADEMIC EXPERIENCE

**Accelerating Graph Neural Network using Heterogeneous System.** 08 2021 - 12.2021
- Explore the heterogeneity of graph structure and propose an ultra-efficient, systolic tensor-based hardware accelerator, with heterogeneous computation paradigm.
- Combined graph clustering and graph neural network to greatly increase data locality and increase the final accuracy.
- Use a heterogeneous system, which includes CPU, FPGA, and AI engine to accelerate computation.

**Memory-Efficient Deep Learning Training via Error-Bounded Lossy Compression** 05.2020 - 08.2020
Training wide and deep neural networks require large amounts of storage resources such as memory.
- Leverages error-bounded lossy compression to significantly reduce the memory requirement for training to allow training larger models or to accelerate training.

**HLS and Verilog Mixed Design for Lossy Compression Algorithm** 03.2020 - 06.2020
- Propose a hardware-algorithm co-design for an efficient and adaptive lossy compressor for scientific data on FPGAs.
- Propose an efficient Huffman coding approach that adaptively updates Huffman codewords online based on our offline Huffman codewords.
- Support a precise control of compression ratio under the error-bounded compression mode.

**Accelerating End-to-End DNN Training via Fine-Grained Pruning** 10.2019 - 02.2020
- Effectively combine the concept of pattern in computer vision with the concept of group lasso in deep neural networks compression.
- Optimize sparse matrix multiplication GPU kernel for sparse convolution and sparse back backpropagation.
- Optimized sparse matrix-matrix multiplication GPU kernel is faster than NVIDIA cuBLAS.

**Geometric Deep Learning based on Directional Curvature for 3D Shape Analysis** 08.2019 - 12.2019
- Geometric deep learning generalizes deep learning models from a 2D Euclidean plane to a 3D geometric surface.
- Proposes a novel geometric deep learning model called CurvaNet that integrates differential geometry with graph neural networks.
- Design a U-Net like architecture with down-/up-sampling paths based on mesh pooling and unpooling operations.

**Algorithm-Hardware Co-Design of 3 Digital Machine Learning ASIC** 08.2017 - 02.2019
- Lead to design and verify three digital chips, which are fabricated with standard 130 nm or 55 nm CMOS technology, and their functionality is correct.
- Design chips adopting multi-core architecture. Users can configure specific cores to participate in computation.
- Implement AXI interface to provide the chip with instruction and high-throughput data.
- Chips are 64× more memory efficient, 51.23× more energy efficient, 10.7× speedup than Intel i5-4200U CPU.

## SKILLS

Programming Language: C, C++, Python, Pascal, MATLAB, Bash
Parallel & Distributed System: MPI, OpenMP, CUDA
Deep Learning Framework: Caffe, TensorFlow, PyTorch

## PUBLICATIONS

- [**ICS'22**] **Chengming Zhang**, et al. "CEAZ: Accelerating Parallel I/O via Hardware-Algorithm Co-Design of Efficient and Adaptive Lossy Compression." arXiv preprint arXiv:2106.13306.
- [**VLDB'22**] Sian Jin, **Chengming Zhang**, Xintong Jiang, Yunhe Feng, Hui Guan, Guanpeng Li, Shuaiwen Leon Song, and Dingwen Tao. "COMET: A Novel Memory-Efficient Deep Learning Training Framework by Using Error-Bounded Lossy Compression" ACM International Conference on Very Large Data Bases, Sydney, Australia, Sep. 5–9, 2022.
- [**ICS'21**] **Chengming Zhang**, Geng Yuan, Wei Niu, Jiannan Tian, Sian Jin, Donglin Zhuang, Yanzhi Wang, Bin Ren, Shuaiwen Leon Song, and Dingwen Tao. "CLICKTRAIN: Enabling Efficient and Accurate End-to-End Deep Learning Training via Fine-Grained Architecture-Preserving Pruning". In Proceedings of the 35th International Conference on Supercomputing (ICS) 2021.

- [**PPoPP'20**] Tian, Jiannan, Sheng Di, **Chengming Zhang**, Xin Liang, Sian Jin, Dazhao Cheng, Dingwen Tao, and Franck Cappello. "waveSZ: a hardware-algorithm co-design of efficient lossy compression for scientific data." In Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP), pp. 74-88. 2020.
- [**DAC'20**] Dong, Peiyan, Siyue Wang, Wei Niu, **Chengming Zhang**, Sheng Lin, Zhengang Li, Yifan Gong, Bin Ren, Xue Lin, and Dingwen Tao. "RTMobile: Beyond Real-Time Mobile Acceleration of RNNs for Speech Recognition." In The 57th Annual Design Automation Conference (DAC). 2020.
- [**KDD'20**] He, Wenchong, Zhe Jiang, **Chengming Zhang**, and Arpan Man Sainju. "CurvaNet: Geometric Deep Learning based on Directional Curvature for 3D Shape Analysis." In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp. 2214-2224. 2020.
- **Zhang, C**. M., G. C. Qiao, S. G. Hu, J. J. Wang, Z. W. Liu, Y. A. Liu, Q. Yu, and Y. Liu. "A versatile neuromorphic system based on simple neuron model." AIP Advances 9, no. 1 (2019): 015324.