Wang Zhang

+86-189-6980-3679 / +1-412-224-3308 zw199006@gmail.com

Education

Carnegie Mellon University, Pittsburgh, PA
Msc. In Mechanical Engineering, December 2015
GPA: 3.85 / 4.0

Zhejiang University, Hangzhou, China
B.Eng. In Applied Mechanics, July 2013
GPA: 3.46 / 4.0

Experience

Senior Software Engineer, Caicloud, Hangzhou, China — 2018 Dec - Present

- Design and develop DL model inference system based on Kubernetes with model optimization, HPA, AB test and feedback logging.
- Design and develop GPU Sharing solution for Kubernetes, sharing one GPU amongst containers with restricted GPU memory and thread usage
- Optimized affinity scheduling for pods of distributed training jobs with PS/Worker mode
- Elastic and fault-tolerant distributed training for data-parallel distributed training (github.com/caicloud/ftlib)

System Engineer II, Netease, Hangzhou, China — 2018 May - Present

- Lightweight environment deployment for deep learning training on Slurm HPC with conda
- Modify Caffe for uneven workload-per-node for distributed training
- Provide acceleration service on distributed training and inference based on TensorRT during Meituan video classification game

Sales Engineer, Wolfram Research, Champaign, IL — 2016 - 2018 April

- Add DL model inference block into Modelica for control
- Hook Modelica into Reinforcement Learning as Plant (environment)
- Prepare an alternative CUDA and OpenCL interface with C-LibrayLink for Mathematica

Research Assistant, PFTLab, CMU, Pittsburgh, PA — 2015-2016

- Implement a CUDA version of CFD-Solver with Immersed Boundary Method and 2nd-order precision, accelerating the fluid simulation speed by 80 times
- Apply viscosity force interpolation for solid body in FSI, improving accuracy in remedy of float limitation
- Re-implement the pressure-based solver with finite volume method with OpenCL and MPI

Skills

Programmings: C++, CUDA, OpenCL, Python, Go, Shell

Github: github.com/zw0610

Blog: zw0610.github.io