Zhenlin Wu

Mobile: +86 18114493674

Email: wuzhenlin456@163.com

EDUCATION

University of Massachusetts Lowell, USA (Scientific Research) **Jan.2018—Jan.2019**

Department of Electrical & Computer Engineering, High-Performance Data Analytics (HPDA) Lab

- ✧ Courses: Computer architecture design, Operating Systems, Software Engineering, Probability and Statistics, High-performance computing on GPUs
- ✧ Positions: Teaching assistant of undergraduate course *Logic Design*
  Research assistant of *High performance graph computing on GPUs*
- ✧ Paper: XBFS: eXploring Runtime Optimizations for Breadth-First Search on GPUs「HPDC19 (CCF B)」

University of California Santa Babara, USA **Sept.2016—Jan.2017**

International Student Exchange Program

Capital Normal University, Beijing **Sept.2014—Jul.2017**

M.S. in Pure Mathematics GPA 84.2/100

Beijing Jiaotong University, Beijing **Sept.2010—Jul.2014**

B.S. in Information and Computing Science GPA 86.0/100

WORK EXPERIENCE

Senior R&D

Baidu, Inc., Shanghai **Jun.2021-present**

- ✧ Participate in the performance optimization of deep learning mobile inference engine Paddle-Lite, and helping to achieve industry-leading performance for different hardware platforms
- ✧ Participate in the design, development and business support of the framework for the deep learning platform Paddle-Lite.

Software development engineer

Suzhou Research Institute, HUAWEI Technologies Co. Ltd **Jul. 2020—May 2021**

- ✧ Participate in the intelligent accelerator group in Collaboration Platform project group and focus on developing and optimizing video processing algorithms for our Audio && video conferencing system product, including video decoder, encoder, CUDA/VIC image resize, composition and OSD caption as well.

Heterogeneous computing engineer

Nanjing Youjia Innovation Technology Co. Ltd (Minieye) **Mar.2019—Mar. 2020**

- ✧ Participate in the development and optimization for company's deep learning inference framework (miniDNN) project in HPC group (using C++ and OpenCL)
- ✧ Also in charge of the maintenance and upgrade of the framework and model performance test.

PROJECT EXPERIENCE

Work project: Baidu, Inc., Shanghai

✧ Concrete work:

a. Focus on OpenCL OP development, functionality enhancement and deep optimization to keep mobile GPU backends performance match to/ ahead of counterparts from competitors; Support more mainstream models to run on mobile-end GPUs;

b. Expand the range of models that OpenCL supports; OpenCL kernel performance enhancements; Actively participate in the online-launching and landing of business models, and solving the feedback problems from the business line.

c. Enhance unit tests support and coverage for the Paddle Inference pass; Resolve inference functionality issues; Promote the stability and usability of the engine.

Work project: Suzhou Research Institute, HUAWEI Technologies Co. Ltd

✧ Participate in the intelligent accelerator group in Collaboration Platform project group

a. Mainly participate in the development of video media processing module in the intelligent accelerator group in our new generation MCU product, which is based on the hardware platform of NVIDIA Jetston AGX Xavier. The Jetson AGX Xavier module has deployed Encoder/Decoder chips to deal with video processing tasks in order to meet the higher requirements of Encode/Decode capabilities or specs in video conferencing products.

b. Mainly responsible for image resize, composition, OSD caption in video processing. CUDA programming is needed owing to the limited composition capability of VIC hardware in Jetson AGX Xavier module. All these units needs to be processed on NVIDIA GPU. High performance resize algorithms implementation in CUDA is required in order to ensure the high quality of images as well as the stability of video processing module in highly concurrent scenarios.

c. Concrete tasks include: utilize CUDA to realize data zero-copy from host to device; register CUDA by utilizing EGL so that the yuv data in block linear format can be directly delivered to the resize component without being transferred into pitch linear format via VIC chip on NV Jetston AGX Xavier module; the resize of block to pitch from decoding unit and the resize of pitch to pitch from yuv data under cross-module transition are both supported.

Work project: Nanjing Youjia Innovation Technology Co., Ltd. (Minieye)

✧ Participate in the development and optimization of the company's self-developed deep learning reasoning framework for mobile device

a. Participate in the development of multi-layer modules of the outer framework (including conv, eltwise and pixelshuffle layers, etc.)

b. Participate in the development of core optimization algorithm related to GPU concerning different GPU hardware features related to Mali and andreno, including F(2x2, 3x3) Winograd algorithm of conv layer and other small layers, in which Winograd algorithm can achieve at least one times the acceleration ratio of the common convolution algorithm of the same scale. The data arrangement adopts NCHWseg8 , which performs best in the company's tests on hardware products. The vector data access of half8 is used in OpenCL to adapt and adjust SIMD operation, which omits the procedure of im2col compared with caffe and matrix multiplication is directly carried out.

c. Responsible for the maintenance and upgrading of the later framework and network performance testing, using GTEST tool for algorithm unit testing.

d. Develop model file conversion script tool proto2json which is converted from Caffe's proto.txt in order to meet the customized requirements of the company's network model

---

Research project: HPDA Lab of ECE Department of UMass Lowell

✧ Parallel Scalable Graph Processing：BFS acceleration on GPUs

a. Responsible for the implementation of the whole project code, using C++ and CUDA in engineering; platform: NVIDIA p6000, TitanXP and V100 GPU
b. Optimization of some uncertain features appearing in parallel BFS algorithm on GPU for the first time
c. Propose the adaptive frontier queue generation strategy to copy with the dynamics exhibited in various stages of BFS traversal, realizing 2.3x average speedup
d. Design two dynamic load balancing strategies to optimize the uncertainty of BFS load in the bottom-up phase, realizing 3.1x average acceleration
e. Introduce an asynchronous strategy for the parallel bottom-up phase, enabling BFS to update multi-level vertices in a round of iterations, and realizing 3.5x average acceleration
f. Overall, 3.5x, 4.9x, 11.2x acceleration of enterprise, TIGR, gunlock and ligra is realized concerning the relevant state-of-the-art projects

SKILLS

Language skills:

✧ TOEFL Test	Total 84; R—27; L—16; S—21; W—20	Jan.2020
✧ GRE Test	Total 313; Verbal—149; Quantitative—164; Writing—3.5	Oct.2021

Technical Skills:

✧ Be familiar with C++ development, data structure and algorithm programming
✧ Have experience in the development of optimization projects of GPU programming such as CUDA and OpenCL
✧ Be familiar with the common Linux system commands and have a basic understanding of operating systems, multi-threading, and computer architecture