Zhihao Li

Curriculum Vitæ

BRIEF BIO

I am currently a Ph.D. candidate at the State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences (CARCH, ICT, CAS) and University of Chinese Academy of Sciences (UCAS), supervised by Prof. Yunquan Zhang. Besides, I am also a joint Ph.D. student in the School of Computational Science and Engineering at Georgia Institute of Technology (CSE, Georgia Tech), supervised by Prof. Richard Vuduc.

Research Interest: high performance computing, optimized FFT library, automatic performance tuning, and many-core/large-scale parallel programming method.

Expected Graduation Date: 2021.

EDUCATION

2015–Now **Ph.D. Candidate**, *Institute of Computing Technology, Chinese Academy of Sciences (ICT, CAS) and University of Chinese Academy of Sciences (UCAS)*, Computer Science, Advisor: Prof. Yunquan Zhang.

2019–Now **Joint Ph.D. Student**, *Georgia Institute of Technology*, Computational Science and Engineering, Advisor: Prof. Richard Vuduc.

2011–2015 **Bachelor of Engineering**, *Guangdong University of Technology*, Computer Science and Technology $(1^{st}/408)$.

HONORS and AWARDS

2019 National Scholarship for Graduate Students.

2019 The 2^{nd} Prize of X86 Track of BenchCouncil International AI System Challenges.

2019 Outstanding Ph.D. Students Scholarship of the Fourth Paradigm Inc.

2018 Joint Ph.D. Student Scholarship of China Scholarship Council.

2018 The First-Class Academic Excellence Scholarship of ICT, CAS.

2018 Merit Student of University of Chinese Academy of Sciences.

2018 Outstanding Student of the State Key Lab. of Computer Architecture, ICT, CAS.

2017 Outstanding Cooperation Award of the Huawei Technologies Co., Ltd.

2017 National Scholarship for Graduate Students.

2017 HPC China 2017, Best Paper Award Finalist.

2016 The First-Class Academic Excellence Scholarship of ICT, CAS.

2016 Merit Student of University of Chinese Academy of Sciences.

2015 Outstanding Undergraduate of University of Chinese Academy of Sciences.

State Key Laboratory of Computer Architecture, ICT, CAS
No.6 Kexueyuan South Road, Zhongguancun, Haidian, 100190 Beijing, China
lizhihao.cs@gmail.com • *Homepage*

PROJECTS

2018–2019 **AutoFFT: a template-based FFT kernel generation framework**.

Facing the challenges such as the increasing complexity and diversity of computer architectures, difficulties in implementing assembly FFT kernels and performance tuning, AutoFFT makes full use of the experience of the domain and optimization experts to automatically generate extremely high-performance FFT code of radices of all natural numbers for ARM, Intel, and AMD processors. AutoFFT thus not only substantially reduces the laborious work of developing assembly FFT kernels manually but also obtains high performance.

- AutoFFT is the first work that systematically summarizes and extracts the integral and general mathematical expressions (the pair and quad patterns) of the symmetric and periodic properties of the DFT matrix. These algorithmic optimizations combines the like terms of the DFT matrix and minimize the amount of floating-point number operations of the FFT butterflies.
- Based on the mathematical expressions, typical and common calculation patterns in the FFT computation are abstracted and defined as templates. These templates are the basis of the template-based framework.
- AutoFFT is the first template-based auto-generation framework, which is able to automatically generate high-performance assembly FFT kernels based on the highly optimized pair and quad patterns for varying CPU architectures.

2017–Now **OpenFFT: A Cross-Platform and High-Performance FFT Library**.

OpenFFT is a portable and high-performance FFT library based on the AutoFFT code generation framework. It can efficiently run on many computing platforms such as Intel, AMD, ARM, Phytium, Hygon and Huawei HiSilicon processors. OpenFFT outperforms the current state-of-the-art FFT libraries.

- OpenFFT generally outperforms the current state-of-the-art commercial and open-source FFT libraries on the aforementioned computing platforms. OpenFFT is averagely 2.14, 1.7, and 2.15 times faster than FFTW, Intel MKL, and ARMPL, respectively.
- OpenFFT has been deployed on Tianhe-3 exascale-class supercomputer prototype. Besides, it can efficiently and stably run on Sugon exascale-class supercomputer prototype and Tianhe-2 supercomputer. Thus, OpenFFT is becoming an important part of the supercomputing ecosystem.
- OpenFFT has been adopted in the Community Earth System Model (CESM) application with an average performance improvement of 12.3% on the FFT-relevant computing modules.
- OpenFFT has been deployed on Huawei ARM-based servers such as Taishan series servers, becoming an important part of the software ecosystem of Huawei Kirin and Kunpeng series processors.

2017–2018 **HartSift: A High-Performance SIFT Implementation on GPUs**.

HartSift is an efficient SIFT implementation on heterogeneous systems (CPU+GPU). It proposes a multi-granularity (thread-warp-block) parallel framework, which constructs dynamic mappings between different granularities and different parallel tasks. HartSift solves the load imbalance with threads during the SIFT feature extraction and realizes real-time feature extraction by making full use of computing resources of CPUs and GPUs.

- The comprehensive performance of HartSift is superior to the current state-of-the-art SIFT implementations. HartSift is 55.88∼121.99 times, 5.17∼6.88 times, and 1.25∼1.79 times faster than OpenCV SIFT, SiftGPU, and CudaSift respectively.
- HartSift adopts two efficient optimizations (rebalancing workloads and multi-granularity parallelism) to address load imbalance among threads (inter-warp workload imbalance and intra-warp workload imbalance) when porting SIFT to GPUs.
- HartSift introduces two high-performance histogram algorithms based on GPUs (the warp-based histogram algorithm and the atomic-free histogram algorithm) to efficiently accumulate information on different scales of samples.

2016–2017 **OpenIPP: A Well-Optimized Library for Multimedia and Data Processing Applications**.

OpenIPP is a well-optimized software library of plenty of functions for multimedia and data processing applications on ARM CPUs. OpenIPP is an equivalent existence on ARM platforms to Intel IPP for diverse Intel processors.

- OpenIPP supports the same functions provided by Intel IPP.
- OpenIPP takes advantage of Single Instruction Multiple Data (SIMD) technique of the ARMv8 architecture to accelerate the supported functions.
- OpenIPP has been deployed on Huawei ARM-based processors such as Huawei Kirin and Kunpeng series processors.

2015–2016 **PerfCV: A Portable and Lightweight Computer Vision library**.

PerfCV is a cross-platform and high-performance computer vision library that supports fundamental image algorithms such as Resize, Cvt_color, GaussianBlur, WarpAffine, WarpPerspective and so on. PerfCV is lighter and faster than OpenCV.

- PerfCV adopts SSE/AVX/NEON intrinsics and Embedded assembly to SIMDize the aforementioned algorithms on x86 and arm CPUs. The CPU version of this library is 1.29∼35.10 times faster than OpenCV.
- PerfCV uses CUDA to implement and optimize the aforementioned algorithms on NVIDIA GPUs. The GPU version of this library is 1.27 2.17 times faster than OpenCV.

PUBLICATIONS

TPDS **Zhihao Li**, Haipeng Jia, Yunquan Zhang, Tun Chen, Liang Yuan, and Richard Vuduc. Automatic Generation of High-Performance FFT Kernels on Arm and x86 CPUs. IEEE Transactions on Parallel and Distributed Systems, 2020. (CCF A; to appear)

SC'19 **Zhihao Li**, Haipeng Jia, Yunquan Zhang, Tun Chen, Liang Yuan, Luning Cao, and Xiao Wang. AutoFFT: A Template-Based FFT Codes Auto-Generation Framework for ARM and X86 CPUs. ACM/IEEE International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC), 2019. (CCF A)

JPDC **Zhihao Li**, Haipeng Jia, Yunquan Zhang, Shice Liu, Shigang Li, Xiao Wang, and Hao Zhang. Efficient Parallel Optimizations of a High-Performance SIFT on GPUs. Journal of Parallel and Distributed Computing, 2019. (CCF B)

CJC Tun Chen, **Zhihao Li***, Haipeng Jia, and Yunquan Zhang. Multi-Dimensional FFT Implementation and Optimization on ARMv8 Platform. Chinese Journal of Computer (In Chinese), 2019. (CCF A; corresponding author)

ICA3PP'18 Xiao Wang, Haipeng Jia, **Zhihao Li***, and Yunquan Zhang. Implementation and Optimization of Multi-dimensional Real FFT on ARMv8 Platform. International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), 2018. (CCF C)

ICPADS'17 **Zhihao Li**, Haipeng Jia, and Yunquan Zhang. HartSift: A High-Accuracy and Real-Time SIFT Based on GPU. IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS), 2017. (CCF C)

PROFESSIONAL EXPERIENCE

2015–2019 **Graduate Research Assistant**, CARCH, Institute of Computing Technology, Chinese Academy of Sciences. Advisor: Prof. Yunquan Zhang.

2017 **Research Intern**, Pengfeng Technologies Co., Ltd. (PerfXLab). Mentors: Dr. Haipeng Jia and Dr. Xianyi Zhang.

2014–2015 **Undergraduate Research Assistant**, CARCH, Institute of Computing Technology, Chinese Academy of Sciences. Advisor: Prof. Yunquan Zhang.

OTHER ACTIVITIES

2020 **Reviewer**, IET Circuits, Devices and Systems(IET CDS).

2019 **Reviewer**, IET Computer Vision(IET CV).

2019 **Presentation**, *AutoFFT: A Template-Based FFT Codes Auto-Generation Framework for ARM and X86 CPUs, (Paper)*, ACM/IEEE International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC'19), Denver, CO, November, 2019.

2017, 2018 **Panels Organizer**, National Annual Conference on High Performance Computing (HPC China).

2017 **Presentation**, *HartSift: A High-Accuracy and Real-Time SIFT Based on GPU, (Paper)*, IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS'17), Shenzhen, Guangdong, December, 2017.

2016 **Volunteer**, National Annual Conference on High Performance Computing (HPC China).