Xiao Wang

Google Scholar Page: https://scholar.google.com/citations?user=MrowKwsAAAAJ&hl=en

Educations

Master of Computer Science, Institute of Computing Technology, Chinese Academy of Science **July 2017-June 2020**

Bachelor of Computer Science, Yunnan University **July 2013-June 2017**

Industry Experience

Senior Software Engineer at Alibaba Cloud Computing Beijing Co. Ltd **July 2020-Now**

Working on optimizing the native c++ SQL engine runtime of MaxCompute (previously named ODPS) which is Alibaba's large-scale computing and storage system.

Key responsibilities:

Optimize Distributed Algorithms

-Design and implement adaptive algorithms to choose suitable physical algorithms(sort, hash) dynamically depending on runtime data characteristics (10%~20% in average for all operator times).

-Design and implement hash-based distributed distinct aggregation (Speedup: 2~50x speedup in average for all jobs).

Optimize Calculations of Aggregation Function

-Develop vectorization implementations for over 30 kinds of aggregate functions to achieve high performance (2x~3x speedup in average for aggregation computing time).

-Develop and optimize computing kernels for no group by key scenario benefiting from SIMD and Loop unrolling (effective for numerical aggregate, 2~3x speedup for computing time).

-Design and implement row-store for partial aggregated states to improve cache efficiency when randomly written on the hash set.

Internship

Software Engineer at Amazon AI Lab(Shanghai) **July. 2019-Aug. 2019**

-Develop and optimize Numpy compatible operators for MXNet

Research Intern at Institute of Computing Technology(Beijing) **May. 2017-Aug. 2017**

-Parallelize and optimize high-performance numeric libraries for ARM CPUs(Cortex-A57) for math primitives (e^x， median filter, etc) by Neon SIMD. This library is compatible with Intel IPP interfaces.

Software Engineer intern at PerfXLab(Beijing) **June. 2018-Aug. 2018**

-Implement and optimize Gaussian Filter for AMD GPU by OpenCL achieving higher performance than this counterpart of OpenCV. This implementation is integrated into the commercial computer vision library provided by PerfXLab.

Research Projects

RealFFT **Sept. 2017-Sept. 2019**

Design a high-performance RealFFT library on ARM CPUs.

This library supports 11 kinds of 1-3 dimensional float/double real FFT algorithms on ARMv8 Arch. Faster than FFTW(http://www.fftw.org/) on ARMv8 platform around 34%~53% for 1D transforms, 10%~41% for 2D transforms, and achieve competitive performances with [MKL](https://software.intel.com/en-us/mkl/features/fft) for some kinds. This work is published in IEEE ICA3PP 2018.

AutoFFT **July. 2018-July. 2019**

AutoFFT is a template-based FFT codes auto-generation framework that contributes to many Chinese vendors' libraries. This work was published in SC19, TPDS 20, and HPCC 21. Contribute to optimizations for specific radix-butterfly kernels on ARM CPUs and all Real FFT implementations.

High-performance SIFT on GPUs [minor contributor]. **July. 2017-Sept. 2017**

Participate in algorithm discussions and some minor modifications.

HartSift is a high-performance SIFT implementation that achieves higher performance than its counterpart in OpenCV, SiftGPU, and CudaSift. This work is published in JPDC in 2019.

Transplant and Performance tuning of OpenBLAS on Tianhe-3 Super Computer (FT2000+ CPU) **Oct. 2018-Nov.2018**

- Performance profiling and tuning for GEMM(SGEMM, DGEMM, CGEMM, ZGEMM)
- Adjust P, Q, R dimensional size and register block size (4x4, 16x4, 8x8) to improve computational intensity.
- Explore and tune instruction order to achieve better CPU pipeline.
- SGEMM, DGEMM, CGEMM and ZGEMM achieve 94.2%, 84.3%, 87.6%, and 95.4% of theoretical peak performance.

MVUC: An Interactive System for Mining and Visualizing Urban Co-locations [author]. **Oct. 2014-July.2014**

Predict and visualize city evolution by mining spatial frequent patterns from Urban data sets. It is published in WAIM 2016.

Virtual machine placement strategy[co-author]. **Oct.2015-July.2016**

This work proposes a cluster-based genetic method to schedule virtual machines more efficiently. It is published in Neurocomputing in 2021.

Publications Lists

-**Implementation and optimization of multi-dimensional real FFT on ARMv8 platform**. International Conference on Algorithms and Architectures for Parallel Processing. Springer, Cham, 2018, **Wang, Xiao**, et al. **[IEEE ICA3PP 2018]**

-**AutoFFT: a template-based FFT codes auto-generation framework for ARM and X86 CPUs**. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-15). Li, Z., Jia, H., Zhang, Y., Chen, T., Yuan, L., Cao, L., & **Wang, X**. (2019, November). **[IEEE SC19]**

-**Efficient parallel optimizations of a high-performance SIFT on GPUs**. Journal of Parallel and Distributed Computing, 124, 78-91. Li, Z., Jia, H., Zhang, Y., Liu, S., Li, S., **Wang, X**., & Zhang, H. (2019). **[JPDC19]**

-**MVUC: An Interactive System for Mining and Visualizing Urban Co-locations**, WAIM 2016 (2): 524-526. **Xiao Wang**, Hongmei Chen*, Qing Xiao, **[WAIM 2016]**

-**Virtual machine placement strategy using cluster-based genetic algorithm**. 2021, 428: 310-316. Zhang B, **Wang X**, Wang H. **[Neurocomputing21]**

Awards & Honors

2019 Merit Student Award of University of Chinese Academy of Science

2018 Second-Class Scholarship of University of Chinese Academy of Science

2016 Elite Collegiate Award. 100 selected junior and senior students in all universities across the country per year

2015 First-Class Scholarship of Yunnan University

2015 Merit Student Award of Yunnan University

2015 Second Prize of Contemporary Undergraduate Mathematical Contest in Modeling

Public Services

-Student Volunteer at HPC CHINA 2018

-Student Volunteer at HPC CHINA 2019

-Serving as volunteer in Mental Health Therapy Center in Kunming.

-Serving as volunteer in water supply for foresters around University of Chinese academy of Sciences.

Technical Skill

Programming language: Experienced with C/C++;Familiar with SIMD, intrinsics; Could use python, OpenCL, SQL

Tools: Git, Linux, CMake, Makefile, GDB, Scons.

OpenSource Project: Familiar with GEMM source Code of OpenBLAS; Familiar with Aggregation Pipe part of Clickhouse.