YU PEI

107 S Mary Ave Apt 87, Sunnyvale, CA, 94086 ◇ (530) 574-7720 ◇ yupei@microsoft.com

SUMMARY

- ◇ **Background in both statistics and computer science**
- ◇ **Can program with C, C++, Python, R and use tools like Git, CMake, Bash**
- ◇ **PhD thesis on optimization of distributed task-based runtime system (PaRSEC) and efficient numerical linear algebra algorithms based on runtime system**

EDUCATION

The University of Tennessee, Knoxville, USA ***Aug. 2016 - Jul. 2022***
Ph.D. Program, Computer Science
Advisors: Jack Dongarra, George Bosilca
Awards: **Graduate Student Senate Travel Awards (Spring, 2020)**

University of California, Davis, USA ***Sep. 2013 - Jun. 2015***
Master of Science, Biostatistics

Sun Yat-Sen University, Guangzhou, China ***Sep. 2008 - Jun. 2013***
Bachelor of Science, Statistics and Biotechnology
Awards: **University Scholarship (2010)**

PROFESSIONAL EXPERIENCE

Software Engineer 2, Azure HPC+AI, Microsoft ***Aug. 2022 ˜Present***

- ◇ **Benchmarking and evaluation of LLM model training and inference on GPU clusters to ensure optimal performance.**

Graduate Research Assistant, University of Tennessee, Knoxville, TN ***Aug. 2016 ˜Aug. 2022***

- ◇ **PaRSEC: Task-based Runtime System, being funded by Exascale Computing Project (ECP); Optimizations of Dynamic task discovery *DTD* interface to enable task graph trimming, and asynchronous broadcast in the runtime.**
- ◇ **DPLASMA: Optimization of dense linear algebra operations for distributed heterogeneous systems using PaRSEC e.g. TRSV, POTRF.**
- ◇ **Stencil Computation with runtime and communication avoidance: Incorporated communication avoiding into task-based runtime implementation to achieve both computation communication overlap and reduction in communication.**
- ◇ **Low-rank and Mix-Precision Cholesky Factorization: Task-based factorization towards Exascale Computing for Climate and Weather Prediction Applications.**

Software Engineering Intern, Cerebras Systems, Sunnyvale, CA ***Summer 2021***

- ◇ **Machine Learning and Math Kernels implementation for the wafer scale engine**

Software Engineering Intern, The Mathworks, Inc, Boston, MA ***Summer 2018***

- ◇ **Worked on the core Simulink engine (C++), enabled multithreaded simulation runs**
- ◇ **Used Pthread, OpenMP and other multithreading libraries**

Data Scientist Intern, Farmers' Business Network Inc, San Carlos, CA ***Summer 2017***

- ◇ **Process propriety farmers harvest data for yield prediction and factor analysis**
- ◇ **Used machine learning algorothms to derive insights for farmer financing and seed procurement**

Graduate student Intern, Oak Ridge National Laboratory, Oak Ridge, TN *2016*

- ◇ **Built in-situ data processing capability into the large scale land simulation model**
- ◇ **Adapted a Fortran parser for automatic code instrumentation**

Graduate Student Researcher, UC Davis, Davis, CA *2014-2015*

- ◇ **Analyzed gridded climate data for biomass growth simulation in the PNW region**
- ◇ **Developed a data processing system using R and Python for regional crops sustainability analysis with HPC system**

PUBLICATIONS

(i) **Qinglei Cao, Yu Pei, Kadir Akbudak, George Bosilca, Hatem Ltaief, David Keyes, and Jack Dongarra. Leveraging parsec runtime support to tackle challenging 3d data-sparse matrix problems. In *2021 IEEE International Parallel and Distributed Processing Symposium (accepted)*, 2021**

(ii) **X. Luo, W. Wu, G. Bosilca, Y. Pei, Q. Cao, T. Patinyasakdikul, D. Zhong, and J. Dongarra. Han: a hierarchical autotuned collective communication framework. In *2020 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 23–34, 2020**

(iii) **Qinglei Cao, Yu Pei, Kadir Akbudak, Aleksandr Mikhalev, George Bosilca, Hatem Ltaief, David Keyes, and Jack Dongarra. Extreme-scale task-based cholesky factorization toward climate and weather prediction applications. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, pages 1–11, 2020**

(iv) **Yu Pei, Qinglei Cao, George Bosilca, Piotr Luszczek, Victor Eijkhout, and Jack Dongarra. Communication avoiding 2d stencil implementations over parsec task-based runtime. In *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 721–729, 2020**

(v) **Qinglei Cao, Yu Pei, Thomas Herauldt, Kadir Akbudak, Aleksandr Mikhalev, George Bosilca, Hatem Ltaief, David Keyes, and Jack Dongarra. Performance analysis of tile low-rank cholesky factorization using parsec instrumentation tools. In *2019 IEEE/ACM International Workshop on Programming and Performance Visualization Tools (ProTools) at SC19*, pages 25–32. IEEE, 2019**

(vi) **Yu Pei, G. Bosilca, I. Yamazaki, A. Ida, and J. Dongarra. Evaluation of programming models to address load imbalance on distributed multi-core cpus: A case study with block low-rank factorization. In *2019 IEEE/ACM Parallel Applications Workshop, Alternatives To MPI (PAW-ATM)*, pages 25–36, 2019**

(vii) **M. Gates, J. Kurzak, P. Luszczek, Yu Pei, and J. Dongarra. Autotuning batch cholesky factorization in cuda with interleaved layout of matrices. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1408–1417, 2017**