Shangyu Luo

lsyurd@gmail.com | 832-270-7372 | Houston, TX | 🏠 Home Page | in LinkedIn

AREAS OF EXPERTISE

Distributed database/data-processing systems and large-scale machine learning

EDUCATION

Rice University 05/2016 – 12/2020

Ph.D. Department of Computer Science

Rice University 09/2012 – 05/2016

M.S. Department of Computer Science

University of Science and Technology of China 09/2008 – 06/2012

B.S. School of Computer Science and Technology

SKILLS

Languages:

- Proficient: C, C++, Java, Python (including NumPy), SQL;
- Familiar: Scala, R, MATLAB, MPI, OpenMP, HTML, Bash, ROS.

Systems:

- Big Data: Hadoop (MapReduce, HDFS, etc.), Spark, SciDB, ScaLAPACK;
- Machine Learning: TensorFlow, Spark MLlib, SystemML.

DevOps Tools: Linux, Docker, Git, SVN, Amazon AWS/EC2/S3, LaTex.

INTERNSHIP EXPERIENCE

IBM - Almaden 06/2015 - 09/2015

- Worked on automatic **operator fusion** to accelerate SystemML, a system for efficiently writing and executing **linear algebra** and **machine learning** computations on **Spark** and **Hadoop**.
- Detected and merged specific linear algebra operators into one compound operator to **reduce** unnecessary **intermediate data** and **computations**.
- **Generated** the runtime code **dynamically** for the compound operators produced by the fusion.
- Conducted a set of preliminary experiments to show that the fused operator could be **up to 300 times faster** than the individual operator for linear algebra computations.

RELATED PROJECTS

Enabling Automatic Matrix Format Exploration for LA and ML 09/2017 – 12/2020

- Proposed a framework to **automatically explore** the formats (e.g., strips or blocks) for vectors and matrices for **linear algebra** (LA) and **machine learning** (ML) computations.
- Formalized the format exploration problem as an **optimization** problem, and solve it using a **dynamic programming** method.
- Designed a **cost model** to estimate the costs for executing LA operations in specific formats.
- Evaluated my ideas on several machine learning and deep learning computations (including a **feed-forward neural network**), and showed that the formats selected by my framework could have a better performance than the ones **manually** picked up by an expert programmer, as well as the ones running on state-of-the-art systems such as **PyTorch** and **SystemML**.

Enhancing Recursion Support for SimSQL 06/2017 – 02/2019
(**2019 VLDB Best Paper Honorable Mention**) (**2019 SIGMOD Research Highlight**)

- Added syntax and function support for complex **recursive algorithms** and **deep learning computations** to SimSQL, an **SQL-based** relational database for large-scale distributed analytics running on **Hadoop** (SimSQL has more than **100,000** lines of code).

- Participated in the design and implementation for the new **query optimizer** for SimSQL to make it capable of optimizing a very large query plan.
- Ran the experiments for the **NLP model Word2Vec** and the **topic model Collapsed LDA**, and showed that the new recursion and optimization support made SimSQL up to **10 times** faster than TensorFlow and Spark for large models with thousands of parameters.

Developing PlinyCompute, a System for Implementing High Performance, Distributed, Data-Intensive Computation 09/2017 – 05/2018

- Participated in the system design and discussion for PlinyCompute (a system with more than **150,000** lines of code).
- Implemented **LDA** and **K-Means** as the first machine learning library for PlinyCompute.
- Realized LDA and K-means on **Spark** with various implementations (RDD and Dataset), and implemented a set of linear algebra computations on **ScaLAPACK**.
- Ran the aforementioned computations on system PlinyCompute, Spark and ScaLAPACK on Amazon EC2 cluster, and had an in-depth comparison for the performances of those systems.

Developing BUDS, a Declarative Language for Machine Learning 02/2015 – 05/2017

- Enhanced BUDS by adding **vector** and **matrix** data types and linear algebra support to it.
- Implemented and ran a few common **Bayesian machine learning** algorithms (e.g., **Bayesian Lasso** and **Hidden Markov Model**) with BUDS and showed that the vector and matrix support accelerated its performance up to **30 times**.

Adding Linear Algebra Support to SimSQL 05/2014 – 08/2017
(**2017 ICDE Best Paper Award**) (**2017 SIGMOD Research Highlight**) (**2020 CACM Research Highlight**)

- Added **vector** and **matrix** as first-class **data types** to SimSQL, and unified its front-end and back-end **type systems**.
- Implemented a set of linear algebra operations and functions for SimSQL.
- Realized **parameterized type signatures** for the functions in SimSQL to make the **optimizer** be aware of the **dimensions** of matrices so that more optimization opportunities are explored.
- Conducted a comprehensive set of experiments to show that the vector/matrix support made SimSQL outperform or have competitive performance with Spark MLlib, SystemML and SciDB on selected linear algebra computations and complex machine learning algorithms.

Predicting Interactions for Proteins and Drugs 09/2014 – 05/2015

- Implemented a **recommendation model** called **Matchbox** using **Markov chain Monte Carlo**.
- Used the MCMC Matchbox model to analyze and predict the protein-protein and protein-drug interactions on the data obtained from Baylor College of Medicine to help the design of drugs.

Comparing ML Algorithms for Distributed Data Analytics Systems 05/2013 - 05/2014

- Implemented a few classic machine learning algorithms (e.g. **Gaussian Mixture Model**, **LDA**, **Hidden Markov Model**, etc.) on **Spark** using **Markov chain Monte Carlo** method.
- Analyzed the strengths and weaknesses of Spark for machine learning applications. Compared Spark with SimSQL, Giraph and GraphLab in regard to performance and usability.

SELECTED PUBLICATIONS (FULL LIST)

Shangyu Luo, Dimitrije Jankov, Binhang Yuan, and Christopher Jermaine. “Automatic Optimization of Matrix Implementations for Distributed Machine Learning and Linear Algebra”. *SIGMOD,* 2021.

Dimitrije Jankov, Binhang Yuan, **Shangyu Luo**, and Christopher Jermaine. “Distributed Numerical and Machine Learning Computations via Two-Phase Execution of Aggregated Join Trees”. *PVLDB*, 14(7):1228-1240, 2021.

Shangyu Luo, Zekai Gao, Michael Gubanov, Luis Perez, Dimitrije Jankov and Christopher Jermaine. “Scalable Linear Algebra on a Relational Database System”*. Commun. ACM*, 63, 8 (August 2020), 93–101. (**2020 CACM Research Highlight**)

Dimitrije Jankov, **Shangyu Luo**, Binhang Yuan, Zhuhua Cai, Jia Zou, Christopher Jermaine, and Zekai J. Gao. “Declarative Recursive Computation on an RDBMS”. *PVLDB*, 12(7): 822-835, 2019. (**Best Paper Honorable Mention**) (**2019 SIGMOD Research Highlight**)

Shangyu Luo, Zekai. J. Gao, Michael Gubanov, Luis Perez, and Christopher Jermaine, “Scalable Linear Algebra on a Relational Database System”. *TKDE*, vol. 31, no. 7, pp. 1224-1238, 1 July 2019. (**Special issue for “Best of ICDE 2017”**)

Jia Zou, R. Matt Barnett, Tania Lorido-Botran, **Shangyu Luo**, Carlos Monroy, Sourav Sikdar, Kia Teymourian, Binhang Yuan, and Christopher Jermaine. “PlinyCompute: A Platform for High-Performance, Distributed, Data-Intensive Tool Development”*. SIGMOD*, pp. 1189-1204, 2018.

Shangyu Luo, Zekai Gao, Michael Gubanov, Luis Perez, and Christopher Jermaine. “Scalable Linear Algebra on a Relational Database System”*. ICDE*, pp. 523-534. IEEE, 2017. (**Best Paper Award**) (**2017 SIGMOD Research Highlight**)

Zekai Gao, **Shangyu Luo**, Luis Perez, and Christopher Jermaine. “The BUDS Language for Distributed Bayesian Machine Learning”*. SIGMOD*, pp. 961-976, 2017.

Tarek Elgamal, **Shangyu Luo**, Matthias Boehm, Alexandre V. Evfimievski, Shirish Tatikonda, Berthold Reinwald, and Prithviraj Sen. “SPOOF: Sum-Product Optimization and Operator Fusion for Large-Scale Machine Learning”. *CIDR*, 2017.

Zhuhua Cai, Zekai Gao, **Shangyu Luo**, Luis Perez, Zografoula Vagena, and Christopher Jermaine. “A Comparison of Platforms for Implementing and Running Very Large Scale Machine Learning Algorithms”. *SIGMOD*, pp. 1371-1382, 2014.

AWARDS AND HONORS

2014/2015 Ken Kennedy-Cray Graduate Fellowship
2017 IEEE ICDE Best Paper Award
2017 SIGMOD Research Highlight
2019 VLDB Best Paper Honorable Mention
2019 SIGMOD Research Highlight
2020 Communication of ACM Research Highlight
Young Researcher Participant for 5th Heidelberg Laureate Forum