Yuanjing Shi

Bellevue, WA 98004

yjshi03@gmail.com shingjan https://shingjan.me

EDUCATION

University of Illinois at Urbana-Champaign **Urbana, IL**
Master of Science in Computer Science GPA: 3.7 / 4.0 *Graduation: May 2020*

The Hong Kong Polytechnic University **Kowloon, Hong Kong SAR**
Bachelor of Science(Hons) in Computing GPA: 3.4 / 4.0 *Graduation: May 2018*

EXPERIENCE

Snowflake Inc. **Bellevue, WA**
Software Engineer *Jun 2020 – Present*

- Built a modern, cloud-built data lake to provide instant elasticity, high availability and on-demand storage with transactional consistency and support for semi-structured data, including JSON, CSV, Parquet, ORC, etc., in the Data Lake and Storage team
- Worked in close cooperation with project managers, support engineers and other functional team members to form a team effort and deliver key features of SnowflakeDB, like Data Export and External Table, for existing and potential customers
- Collaborated with other members among different organizations and teams with extensive experience in large-scale, distributed system architectures and agile development environment

Microsoft **Beijing, China**
Software Engineering Intern *May 2018 - Aug 2018*

- Designed and developed the transactional branch of GraphView, a middleware and DLL library for the Graph Gremlin API of Microsoft's Azure CosmosDB, in collaboration with the Intelligent Cloud & Edge Group and the Azure team
- Constructed the testing framework based on YCSB & TPC-C in order to benchmark GraphView against other distributed databases on multi-node clusters

PROJECTS

SnowflakeDB - Data Lake *Jun 2020 - Present*

- Pushed Data Export feature from Private Review to Public Review by designing and implementing key functionalities to unload semi-structured data to cloud storage with metadata management and partitioning
- Engaged in the process of delivering External Table of SnowflakeDB to General Availability (GA) with design reviews, customer issue investigations and regression testing
- Improved the stability and test coverage of Snowflake's Data Lake by integrating native data generator into regression tests
- Built a heat map for External Table for existing customers to significantly reduce storage cost at Snowflake's annual hackathon

LLVM-based Machine Learning Compiler for ONNX *Aug 2019 - Jun 2020*

- Built a compiler infrastructure to map high-level deep learning models to more specific and performance-driven C/C++ programs for different runtime systems
- Designed and constructed the front-end which takes a structural description and converts it to LLVM/HPVM-supported C/C++ code
- Extended the Nvidia-cuDNN-based back-end to generate low-level primitives for various hardware architectures and enable various optimization strategies, like operator confusion, for model inference

Flare - Native Compiler Framework for Apache Spark *Feb - Aug 2019*

- Built Flare, an efficient Spark-like data processing framework, with native compilation on its Scala/SQL front-end, which outperforms Apache Spark by up tp 10 times on standalone and computation-intensive workloads like TPC-H
- Integrated Message Passing Interface (MPI) into Flare's compiler framework to support distributed data processing
- Delayered Hadoop HDFS with Direct I/O access via zero-copy/memory map to avoid data duplication

Leyenda - an Adaptive, Hybrid Radix Sorting Algorithm *Jan - May 2019*

- Designed and implemented Leyenda to efficiently sort large-scale data both internally and externally with adoptions of Producer-Consumer I/O access pattern, memory map, PageCache and sorting based on radix-bit (60 GB in 290 seconds)
- Competed in the ACM SIGMOD 2019 Programming Contest and placed Leyenda as one of the five finalists among 35 other teams worldwide (4 out of 35), with one travel grant awarded to present our design and implementation at SIGMOD' 19 in Amsterdam

Azure CosmosDB - Distributed NewSQL Database *May - Aug 2018*

- Finalized the concurrency protocol implemented in VersionDB, GraphView's transaction branch, based on a MIT-proposed novel optimistic concurrency control algorithm called TicToc
- Constructed benchmark framework for the Cassandra-based VersionDB and conducted benchmark testing against competitor's offerings like CockroachDB and DBX1000 on multi-node cluster with YCSB/TPCC workloads and HAproxy for load balancing

SKILLS

- Programming Languages:
  - Over 20000 lines: Java • C/C++ • Python • Scala
  - Over 5000 lines: Go • Rust • C# • OCaml • Matlab
- Data Intensive Computing: Apache Arrow • Hadoop • Parquet • Spark • HDFS
- Deep Learning Frameworks: TensorFlow/Keras • PyTorch • ONNX • Apache TVM
- Cloud Native Computing: Linux • Docker • Kubernetes • Load-balancing