QING XIAO

+1 773-382-9896 | Redmond, WA | xqoasis@gmail.com | linkedin | github

Education

University of Chicago Sep. 2022 – Mar. 2024
*M.S. in Computer Science (**High-Performance Computing** track)* *Chicago, IL*

Shanghai Jiao Tong University Sep. 2018 – Jun. 2022
B.A. in Japanese, Minor in Applied Mathematics *Shanghai, China*

Experience

Software Engineer | Microsoft | Infra, Confidential Computing May. 2024 – Present | Redmond, WA

- Led an **AI Classification** project (Microsoft Hackathon 2024), improving Azure Copilot query accuracy.
  - Deployed a fine-tuned **Small Language Model** to predict Azure Resources from user queries with probabilities.
  - Implemented **adaptive RAG** based on SLM generated probabilities and tenant resource data, improving the accuracy of the query to 90%.
  - Developed in **KAITO** (Kubernetes AI Toolchain Operator) for efficient utilization of GPU resources. Integrated with **Azure OpenAI** for real-time query processing.
- **Independently designed and developed** a scalable **E2E cloud-native** VM image validation service. Restores metadata in the message queue, with a message bus to pull out and send authenticated request to Azure Resource Manager, and on-prem processes Vhd files asynchronously. It includes **multi-authentication protocols**, **platform attestation** mechanisms, certification rotation through Azure.
- Implemented **Confidential Computing VM** (CVM) attestation extension for **AMD SEV-SNP** (Secure Encrypted Virtualization-Secure Nested Paging) CPU. Deployed distributed attestation validation and reporting **micro-service** through Azure Kubernetes Service **(AKS)**, achieving high availability across production regions.
- Enhanced **Infrastructure Guest Virtual Machine Agent** (IgvmAgent) self-signed certification rotation. Also configured the deployment action plan for Azure Stack Hyperconverged Infrastructure (Azure Local OS).
- Engineered E2E **CI/CD** infrastructure, spanning assembly build, **on-prem VM provisioning**, extension deployment, and attestation across Azure regions. Added pipeline fall-back scenario and validated stable receipt to increase pipeline reliability **50% to 100%**.

Research Assistant | University of Chicago | System Lab Sep. 2023 – Mar. 2024 | Chicago, IL

- Researched cross-platform performance portability in **heterogeneous runtime systems**, focusing on graph applications using the **Gunrock** lib. Developed machine learning models (**CNN** and **RF**) to predict key GPU performance metrics (IPC, warp execution efficiency, DRAM read throughput) across Nvidia **A100/V100/P100.**
- Implemented feature selection using mutual information, data preprocessing with **MinMaxScaler**, and performance profiling using CUDA tools (**nvprof** and **nsys**).
- Built an automated workload distribution system with **Slurm** for scalable model training and evaluation.

Software Engineer Intern | Oracle | Java Core Lib Jun. 2023 – Sep. 2023 | Santa Clara, CA

- Contributed to Java (**OpenJDK 22**)'s new **Classfile API** for bytecode manipulation, replacing **ASM** library.
- Created 250+ functional tests for **Java Core Lib**, **Javac Compiler**, **Language Tools** to create modern **byte-code** level operation tests. All code was integrated into OpenJDK mainline and was nominated as **OpenJDK Author**.
- Developed a VS Code IDE **Java extension** that features code refactoring, smart editing, JDK 21 support, etc. Implemented Language Server based on NetBeans **Language Server Protocol** (LSP).
- Optimized **JLink Plugin** to speed up the creation of **Modular Java Run-Time Images** and enhanced Java class constant pool's traversing method.

Software Engineer Intern | DiDi | Self-driving Mar. 2022 – Jul. 2022 | Shanghai, China

- Developed user interfaces with Qt to System on Chip (**SoC**), encompassing five data monitoring functions.
- Built **cross-compilation** solution through CMake and Makefile to reduce compile time.

Data Science Intern | Eli Lilly and Company Jul. 2021 – Dec. 2021 | Shanghai, China

- Developed automated **A-B test tool**. Designed **Naive Bayes-based opinion clustering system** (83% accuracy)

Projects

AWS-hosted Hadoop-based Distributed System | *Hadoop, Spark, AWS* Apr. 2023 – Jun. 2023

- Developed a scalable big data analysis system on **AWS EMR/EC2** using Lambda Architecture (**Hive/HDFS** batch, **Spark/Kafka** streaming). Enhanced scalability via **S3 URL, SNS, SQS, DynamoDB** (800k+ RPs/instance)

Technical Skills

HPC, AI: GPU, CPU; PyTorch, TensorFlow; CUDA, OpenMP, Triton; SGLang, vLLM;
Infra, Cloud, Language: Cpp, Python, Go, Java, SQL; Kubernetes, Docker, Redis, Slurm; Azure, GCP, AWS