Chi Zhang

Resume

Personal Inforamtion

Name Chi Zhang

Phone 757-338-7196

Email chzh.xdp@gmail.com

ShortBio Active researcher and software engineer. Experienced and interested in the areas of **system, infrastructure, deep learning and Generative AI.**

Education

Ph.D. in Computer Science, *University of Pittsburgh*, 2022.

B.E. in Computer Science, *Xidian University*, China, 2014.

Work Experience

AMD (Advanced Micro Devices)

MTS (Member of Technical Staff) Software Engineer, San Jose, CA.
AI System & Runtime, 2023.05 - Current
[NPU, LLM, Stable Diffusion]

- Enhanced inference performance for large language models (LLMs) and generative AI on the AMD RyzenAI Neural Processing Unit (NPU), the world first dedicated AI engine integrated into an x86 CPU.
  - Enabled efficient execution of open-source LLMs (e.g., Llama, Google Gemma, Alibaba Qwen, DeepSeek) and various Stable Diffusion models (including SD1.4, SD2.1, SD3, and SDXL-Turbo) by fully leveraging the specialized architecture of the RyzenAI NPU.
- Collaborated with cross-functional teams to design and integrate custom ops kernels and specialized compilation infrastructure for the NPU. Strategically offload compute-intensive operations such as basic ops like Conv/Gemm and complex ops like MHA, MLA, GQA, and Rope to the NPU, ensuring a performant and efficient bringup of next-generation generative AI models.
- Gained in-depth expertise in the latest breakthroughs in LLM and Stable Diffusion technologies through hands-on deployment of cutting-edge generative AI models.
- Participated in the development of a unified deep learning inference server, facilitating containerized model uploads and real-time inference across heterogeneous computing backends (CPU, GPU, and FPGA).

Google ***Software Engineer (L4)***, Sunnyvale, CA.

Payments Invoicing, 2020.09 - 2023.04

Support the Invoicing of Google

[Java, SQL, ProtoBuf, Flume (in-house Apache Beam), Borg (in-house Kubernetes)]

- Participated in the APIs or services design/implementation/testing for the purpose of generating/maintaining/modifying the invoices/documents/transactions for all Google customers.
- The services and APIs we are building are targeted to be planet-scale, real time and stable. The APIs are deployed on the Google's robust payment platform which ensures the performance and scalability.

Software Engineer PhD Intern, Mountain View, CA.

Google Payment Infrastructure, 2017.05-08

Support the Invoicing of Google [Java, SQL, ProtoBuf]

- Participated in the specific project to help migrate the BigTable/MapReduce based backend service to use a new Google F1 based system.

Meta ***Software Engineer PhD Intern***, Menlo Park, CA.

PyTorch Glow Runtime team, 2018.06-08

Quantization Support for GPU backend of Glow [C++, OpenCL]

- Add quantization support for more than 20 GPU operators of Glow.
- Enable the weights and data of the neural network to be stored in quantized format (INT8) other than 32bits (INT32). **Reduced the entire memory usage by 75%.**

Software Engineer PhD Intern, Menlo Park, CA.

PyTorch Glow Runtime team, 2019.05-07

(Glow: A machine learning compiler and execution engine for hardware accelerators. The compiler is designed to allow state of the art compiler optimizations and code generation of neural network graphs.)

Support for debugging in Glow [C++, Python]

- Add functionality to track and dump all changes that happened in the graph compilation and optimization phases of Glow.
- Implemented log-based debugging tools to
  - reconstruct the node graph at any certain fixed compilation phase of Glow.
  - filter and infer all nodes transformations related to one given node.
  - collect basic statistics of nodes at any phase or bewtween any pair of compilation phases.

Bosch Research & Technology Center ***Research Engineer Intern***, Pittsburgh, PA.

Privacy and Security team, 2016.05-07

Cloud-based encrypted search engine [Java, Apache Lucene/Solr]

- Designed and implemented an encrypted search engine infrastructure that is based on SSE (Searchable Symmetric Encryption).
- Achieve scalability for this infrastructure by utilizing Apacha Lucene/Solr and deployment on AWS.

Publications

Exploiting the Regular Structure of Modern Quantum Architectures for Compiling and Optimizing Programs with Permutable Operators, Yuwei Jin, Fei Hua, Yanhao Chen, Ari Hayes, **Chi Zhang**, Eddy Z. Zhang . The 28th International Conference on Architectural Support for Programming Languages and Operating Systems **(ASPLOS 2023)**, Vancouver, Mar 2023. .

Time-Optimal Qubit Mapping, **C. Zhang**, A. Hayes, L. Qiu, Y. Jin, Y. Chen, E.Z. Zhang. The 26th International Conference on Architectural Support for Programming Languages and Operating Systems **(ASPLOS 2021)**, Virtual, Apr 2021.

AutoBraid: A Framework for Enabling Efficient Surface Communication in Quantum Computing, F. Hua, Y. Chen, Y. Jin, **C. Zhang**, A.B. Hayes, Y. Zhang, E.Z. Zhang. The 54th IEEE/ACM International Symposium on Microarchitecture **(MICRO 2021))**, Virtual, Oct 2021.

Locality-Aware Software Throttling for Sparse Matrix Operation on GPUs, Y. Chen, A. Hayes, **C. Zhang**, T. Salmon, E.Z. Zhang. Proceedings of the USENIX Annual Technical Conference **(USENIX ATC 2018)**, Boston, MA, July 2018.

Live Code Update for IoT Devices in Energy Harvesting Environments, **C. Zhang**, W.Ahn, Y.Zhang, B.Childers. Non-Volatile Memory Systems and Applications Symposium **(NVMSA)**, 2016 5th. IEEE, 2016.

Skills

Languages	C++/C, Python, Java
Tools	PyTorch, ONNX Runtime, ONNX
AI Knowledge	Transformers, Stable Diffusion
CS Knowledge	Solid algorithms skills, Computer Systems concepts (e.g. I/O system, compiler, distributed system)

Awards

National Lizhi Scholarship, Ministry of Education, China

National Scholarship, Ministry of Education, China

Services

Reviewer and Program Committee Member CSCWD 2024 (2024 27th International Conference on Computer Supported Cooperative Work in Design)

Reviewer International Symposium on Code Generation and Optimization (CGO) 2024

Reviewer ACM Transactions on Embedded Computing Systems 2020

Reviewer IEEE Internet Computing 2021

Artifact Evaluation Committee Member PPoPP 2019

Artifact Evaluation Committee Member CGO-PPoPP 2017