Hao (Mark) Chen

+44 (0) 7419584455
hc1620@ic.ac.uk
github.com/hmarkc
linkedin.com/in/mark-hao-chen

EDUCATION

Imperial College London
PhD in Computing *Sep 2024 – June 2028 (expected)*

- **Research Focus**: High-Level Synthesis, High Performance Computing, Hardware Acceleration, Machine Learning System, Large Language Model
- **Awards**: President's PhD scholarships (2024-2028)

Imperial College London
MEng in Computing *Oct 2020 – June 2024*

- **CS Modules**: Custom Computing, Advanced Computer Architecture, Computer Vision, Robotics, Operating System, Algorithm Design and Analysis, Compiler, Network and Communication
- **Mathematics Modules**: Linear Algebra, Probability and Statistics, Operation Research
- **GPA**: 87.40% (Year 1), 82.62% (Year 2), 84.60% (Year 3)
- **Awards**: Dean's List (Year 1-3); G-Research Ltd Prize, Aug 2021

RESEARCH

Parallel Prompt Decoding: Hardware-Aware Memory-Efficient Acceleration of LLM Inference
Imperial College London *Jan 2024 – May 2024*

- Developed Parallel Prompt Decoding (PPD) to accelerate LLM inference, training on a single A100-40GB GPU in 16 hours with only 0.0002% trainable parameters.
- Achieved up to 2.49× speedup and maintained minimal runtime memory overhead of 0.0004% on LLMs from MobileLlama to Vicuna-13B, across benchmarks like MT-Bench and HumanEval.
- Proposed a hardware-aware dynamic sparse tree technique, optimizing PPD's performance for each specific hardware platform.
- Open-sourced implementation of PPD, facilitating broader adoption and collaboration: repository url.

AutoBayes: Fast Uncertainty Estimation using Bayesian Neural Network on FPGA [2]
Imperial College London *July 2022 – Aug 2023*

- Built an automatic tool to transform traditional Neural Networks to Bayesian Neural Networks (BNNs) using Monte-Carlo Dropout (MCD) in Keras framework; extended the tool hls4ml to generate fast and power-efficient Bayesian hardware designs for Xilinx FPGAs from BNNs
- Developed a transformation framework involving four phases for multi-exit MCD-based BNNs: optimizing architecture, spatial and temporal mapping optimization, algorithm-hardware co-design, and HLS-based hardware accelerator generation; this framework systematically and effectively explores the design space of multi-exit MCD-based BNNs for their efficient implementation
- Implemented multi-exit mask-based BNN transformation inspired by Masksembles, to enhance the multi-exit MCD-based BNN approach; utilizing pre-defined dropout masks on a shared single neural network to reduce memory overhead as compared to deep ensembles, and controlled overlap and correlation among masks to achieve similar algorithmic performance as traditional deep ensembles

Deep QLearning Scheduler to Enhance Task Placement in Fog Computing
Imperial College London *March 2023 – June 2023*

- Implemented the Deep QLearning Scheduler algorithm for container migration in Fog Computing (FC) environments, proposed by the paper *Migration Modeling and Learning Algorithms for Containers in Fog Computing*
- Integrated the Deep QLearning algorithm into the COSCO (Container Orchestration Using Co-Simulation and Gradient Based Optimization for Fog Computing Environments) framework, enabling intelligent task placement and management in large-scale fog platforms; used the simulator to obtain environmental rewards and make migration decisions

PUBLICATIONS

[1] Zehuan Zhang, Hongxiang Fan, **Hao (Mark) Chen**, Lukasz Dudziak and Wayne Luk. Hardware-Aware Neural Dropout Search for Reliable Uncertainty Prediction on FPGA. *2024 Design Automation Conference (DAC).*

[2] Hongxiang Fan, **Hao (Mark) Chen**, Liam Castelli, Zhiqiang Que, He Li, Kenneth Long, Wayne Luk. When Monte-Carlo Dropout Meets Multi-Exit: Optimizing Bayesian Neural Networks on FPGA. *2023 Design Automation Conference (DAC).*

[3] **Hao (Mark) Chen**, Taowen Liu, Songyun Hu, Leyang Yu, Yiqi Li, Sihan Tao, Jacqueline Lee, Ahmed E. Fetit. Web-based AI System for Medical Image Segmentation. *2023 Conference on Medical Image Understanding and Analysis (MIUA).*

SUBMITTED MANUSCRIPTS

- **Hao (Mark) Chen**, Wayne Luk, Ka Fai Cedric Yiu, Rui Li, Konstantin Mishchenko, Stylianos I. Venieris, Hongxiang Fan. Hardware-Aware Parallel Prompt Decoding for Memory-Efficient Acceleration of LLM Inference. Submitted to *2024 Conference on Neural Information Processing Systems (NeurIPS).*
- **Hao (Mark) Chen**, Liam Castelli, Martin Ferianc, Shuanglong Liu, Wayne Luk, Hongxiang Fan. Algorithm and Hardware Co-Design for Multi-Exit Dropout-based Bayesian Neural Networks. Sumbitted to *2024 IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I).*

INDUSTRIAL EXPERIENCE

Qube RT

Quantative Technologist Intern, UK *April 2023 – Sep 2023*

- Developed a C++ monitoring system for thread pool performance, utilizing the blink protocol for data serialization and publishing to other services; integrated Prometheus Database and Grafana to visualize and analyze performance statistics
- Created a service within a low-latency trading platform responsible for aggregating performance statistics and publishing them at regular intervals; achieved persistence of the statistics by utilizing ODB (Object-Relational Mapping library) and PostgreSql database

Huawei Technologies Research & Development

Graphics Modelling Intern, UK *March 2022 – Sep 2022*

- Built an application using Jinja template engine to deserialize specification in xml and json format to C++ structures, functions, and definitions as part of the graphics API; completed a profile generator to produce valid graphics API profiles from given schema in json format using Python
- Wrote Python scripts to convert between xml and json files used for the API specification; used Flatbuffers to drive glTF sample generation efficiently

Ampere Computing Shanghai, China

Java Software Developer *June 2021 – Sep 2021*

- Developed open-source plugins for Jenkins, a leading CI/CD platform, using JAVA/JELLY; became the maintainer of Lucene Search Plugin, an open search tool plugin; fixed Out Of Memory Exception of Lucene Search Plugin when handling over 100 GB of data
- Improved the indexing speed of Lucene Search Plugin by more than 50% after structure optimization; enriched the searching option and added pagination

SKILLS

- **Programming:** C++, C, Python, Scala, Java, Swift, Haskell, Bash
- **Tools:** GCC, Jenkins, Github, Docker, Heroku, AWS
- **Framework:** PyTorch, Keras, ODB, Kitura, Lucene, Jinja2
- **Language:** GRE - 333/340 + 5/6 (Verbal 163, Quantitative 170, Analytical Writing 5); TOEFL iBT 113/120 (Reading 30, Listening 29, Speaking 26, Writing 28)