Yaoyao Ding

yaoyao@cs.toronto.edu linkedin.com/in/yyding cs.toronto.edu/~yaoyao github.com/yaoyaoding

Education

University of Toronto **Toronto, Canada**
PhD Student, Department of Computer Science *Sep, 2022 - current*
Supervisor: Gennady Pekhimenko

University of Toronto **Toronto, Canada**
MASc in Computer Engineering, Department of Electrical and Computer Engineering *Sep, 2020 - June, 2022*
Supervisor: Gennady Pekhimenko
Thesis: "IOS: Inter-Operator Scheduler for CNN Acceleration"

Shanghai Jiao Tong University **Shanghai, China**
Bachelor in Computer Science, School of Electronic Information and Electrical Engineering *Sep, 2016 - June, 2020*
Program: Zhiyuan Honor Program (ACM Class)

Publications and Manuscripts

- Hexcute: A Tile-based Programming Language with Automatic Layout and Task-Mapping Synthesis.
  Xiao Zhang, Yaoyao Ding, Yang Hu, Gennady Pekhimenko
  manuscript
- Tilus: A Tile-Level GPGPU Programming Language for Low-Precision Computation
  Yaoyao Ding, Bohan Hou, Xiao Zhang, Allan Lin, Tianqi Chen, Cody Yu Hao, Yida Wang, Gennady Pekhimenko
  to appear in ASPLOS 2026
- Grape: Practical and Efficient Graphed Execution for Dynamic Deep Neural Networks on GPUs.
  Bojian Zheng, Cody Hao Yu, Jie Wang, Yaoyao Ding, Yizhi Liu, Yida Wang, Gennady Pekhimenko
  Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2023)
- Hidet: Task Mapping Programming Paradigm for Deep Learning Tensor Programs.
  Yaoyao Ding, Cody Yu, Bojian Zheng, Yizhi Liu, Yida Wang, and Gennady Pekhimenko
  Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS 2023)
- IOS: Inter-Operator Scheduler for CNN Acceleration.
  Yaoyao Ding, Ligeng Zhu, Zhihao Jia, Gennady Pekhimenko, Song Han
  Proceedings of Machine Learning and Systems, Volume 3 (MLSys 2021)
- GAN Compression: Efficient Architectures for Interactive Conditional GANs.
  Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, Song Han
  Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)
- Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario.
  Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, Zhenhui Li
  Proceedings of The Web Conference 2019 (WWW 2019 Demo)

Experience

NVIDIA **Toronto, Canada**
Software Engineer *June, 2025 - current*
- CentML was acquired by NVIDIA.
- Open-source Tilus, a tile-level GPU kernel programming language: https://github.com/NVIDIA/tilus.

CentML **Toronto, Canada**
Research Engineer (Founding Engineer) *July, 2022 - June, 2025*
- I joined the startup founded by my supervisor, conducting research and development of deep learning compiler.
- Open-source Hidet: https://github.com/hidet-org/hidet.
- Lead the development of open-source project Hidet: 1) support dynamic shape; 2) add Hidet Script, allowing developers to develop kernels in Python in a simple way; 3) add hidet as a backend for torch dynamo (e.g., torch.compile); 4) develop a new tile-level GPU programming language (tilus).

Amazon Web Services (AWS) **Santa Clara, USA**

Applied Scientist Intern *Dec, 2021 - Apr, 2022*

- Design and implemented a new deep learning compiler, Hidet, from scratch to address the limited expressive ability of existing state-of-the-art DNN compiler TVM.
- Hidet outperforms the best of PyTorch, ONNX-Runtime, AutoTVM, and Ansor on five representative models (ResNet50, Inception V3, MobileNetV2, Bert, GPT-2) with up to 1.48x (1.22x on average) speedup. This work has published at ASPLOS '23.

Amazon Web Services (AWS) **Shanghai, China**

Applied Scientist Intern *Mar, 2021 - Jun, 2021*

- I added inter-operator parallelization support for a TVM-based machine learning framework. More specifically, new virtual machine instructions and relay operators are added to allow us to control the CUDA stream each operator launched on.
- Besides, three schedulers are implemented to schedule the inter-operator parallelization. Inter-operator parallel execution achieves up to 1.25x speedup.

HAN Lab, Massachusetts Institute of Technology (MIT) **Cambridge, USA**

Research Assistant *Jul, 2019 - Dec, 2019*

- We pointed out a bottleneck for efficient CNN inference: existing intra-operator parallelism can not saturate the high parallelism of modern hardware, especially for recent CNN models with more smaller convolutions.
- I proposed a novel dynamic programming algorithm to find a schedule for inter-operator parallelization. The schedules IOS generated consistently outperform existing deep learning libraries (e.g. TensorRT) by 1.1-1.5$\times$ in terms of latency. This work has been published at MLSys '21.
- During the internship, I also helped my college students to deploy the compressed GAN models to edge devices, such as Raspberry Pi, NVIDIA Jetson Nano, and mobile phones using TVM. This work has been published at CVPR '20

Apex Lab, Shanghai Jiao Tong University (SJTU) **Shanghai, China**

Research Assistant *Sep, 2018 - Jun, 2019*

- I and my colleagues designed and implemented a new traffic simulator CityFlow with high effeciency. Cityflow is more than twenty times faster than baseline SUMO and is capable of supporting city-wide traffic simulation with an interactive render for monitoring.
- We also provided a user-friendly interface for reinforcement learning in Python package. This work has been published at WWW 2020 as a demo paper.

Honors and Awards

- Qualcomm Innovative Fellowship ($50,000) *2023*
- Amazon Post-Internship Fellowship ($10,000) *2022*
- Liao Kaiyuan Scholarship (20,000 CNY) *2019*
- Rong Chang Innovation Scholarship Finalist (10,000 CNY) *2018*
- Zhiyuan Honorary Scholarship (10,000 CNY per year) *2017, 2018*
- Gold Medal in ACM-ICPC Qingdao Asia Regional Contest (top 5% over 80 teams) *2017*
- Gold Medal in CCPC Harbin Regional Contest (top 6% over 180 teams) *2017*
- Gold Medal in CCPC Hefei Regional Contest (top 12% over 150 teams) *2016*
- Bronze Medal in Chinese National Olympiad in Informatics *2015*

Academic Services

- Program Committee: NeurIPS (2023, 2025), ICML (2024, 2025), ICLR 2025.
- Artifact Evaluation Committee: MLSys 2023.
- Organizer of tutorial "Predicting and Optimizing Runtime Performance of Deep Learning Models" at ASPLOS '23.

Talks

- Accelerating DNN Inference with End-to-End Compilation — Santa Calra, USA
  NVIDIA GTC '24. *March 2024*
- Hidet: Task Mapping Programming Paradigm for Deep Learning Tensor Programs — Online
  Google Lab. *June 2023*
- Hidet: Task Mapping Programming Paradigm for Deep Learning Tensor Programs — Vancouver, Canada
  ASPLOS '23. *March 2023*
- Build Tensor Programs with Hidet in Python — Vancouver, Canada
  ASPLOS '23 Tutorial. *March 2023*
- Hidet: Task Mapping Programming Paradigm for Deep Learning Tensor Programs — Online
  TVM Conference 2023. *March 2023*
- Multi-stream Support for Virtual Machine Executor — Online
  TVM Conference 2021. *Dec. 2023*
- IOS: Inter-Operator Scheduler for CNN Acceleration — Online
  MLSys '21. *April 2021*

Technical Skills

Languages: C/C++, Python, CUDA, PTX, Java
Developer Tools: Git, VIM, Nsight Compute/Systems, docker
Technologies/Frameworks: TVM, cuBLAS, cuDNN, OpenAI Triton, MLIR, LLVM, vLLM, TensorRT-LLM

Update: September, 2025