Abdul Dakkak

Compiler Developer · System Architect · AI Optimizer

2906 S. Myra Ridge Dr, Urbana, IL 61802

(+1) 419-418-1158 | dakkak@illinois.edu | www.dakkak.dev | abduld

Research Interest

My research interest lies between programming languages and accelerated computing. My work has focused on understanding and optimizing high-level languages that target accelerators. In the process, I have developed widely used industry-grade tools for compiling, running, profiling, and introspecting whole-level applications to optimize their performance across both the hardware and software stack.

Education

PhD. in Computer Science — *Champaign, IL*

University of Illinois at Urbana-Champaign — *Aug. 2013 - Exp. Aug. 2020*

B.A. in Pure Mathematics — *Toledo, OH*

University of Toledo — *Aug. 2005 - June. 2009*

Experience

Senior Compiler Developer – Wolfram Research — *Jan. 2019 - Present*

- Continued to architect and develop the Wolfram Language compiler.
- Architected the runtime library used for the Wolfram language.
- Developed a constraint-based type system for the Wolfram language.
- Developed datastructures and applications leveraging the compiler (in *Mathematica* 12+)

Kernel Developer – Wolfram Research — *April. 2010 - Dec. 2018*

- Architected the Wolfram compiler which was released in *Mathematica* 12.
- Developed a domain specific language to write financial code for Wolfram Finance Platform (released in *Mathematica* 11).
- Developed graphics rendering for the cloud using both canvas and WebGL (released in *Mathematica* 9).
- Developed optimized data-structures and algorithms for computational geometry (released in *Mathematica* 10).

Junior Kernel Developer – Wolfram Research — *April. 2009 - April. 2010*

- Developed Wolfram's CUDA and OpenCL integration (released in *Mathematica* 8).
- Enhanced the C/C++ library bindings interface of the Wolfram language.

Projects

Team Lead & Senior Compiler Developer – Mathematica (wolfram.com/mathematica) — *2009 – Present*

- Developed the Wolfram type system, runtime, and compiler.
- Optimized core Wolfram engine for desktop and cloud.
- Developed CUDALink and OpenCLLink.
- Developed a DSL for writing GPU and CPU financial indicators.

System Architect & Primiary Developer – MLModelScope (github.com/rai-project/mlmodelscope) — *2017 – 2020*

- An inference system designed for hierarchical profiling and benchmarking.
- Over 300 built-in models supported and integration with MLPerf workloads.

System Architect & Primiary Developer – RAI (github.com/rai-project/rai) — *2016 – 2019*

- A system designed as a configurable programming environment for heterogeneous parallel programming.
- An interactive command line tool used for building and executing accelerated code in the cloud.

System Architect & Primiary Developer – WebGPU (webgpu.com) — *2013 – 2018*

- A lab-submission system designed for developing CUDA and OpenCL programming within the browser.
- Used by over 100,000 students to evaluate millions of labs.

Publications

DLSpec: A Deep Learning Task Exchange Specification — *USENIX OpML 2020*
Adbul Dakkak, Cheng Li, Jinjun Xiong, Wen-mei Hwu

Benanza: Automatic μBenchmark Generation to Compute "Lower-bound" Latency and Inform Optimizations of Deep Learning Models on GPUs — *IPDPS 2020*
Adbul Dakkak, Cheng Li, Jinjun Xiong, Wen-mei Hwu

DLBricks: Composable Benchmark Generation to Reduce Deep Learning Benchmarking Effort on CPUs — *ICPE 2020*
Cheng Li, **Adbul Dakkak**, Jinjun Xiong, Wen-mei Hwu

XSP: Across-Stack Profiling and Analysis of Machine Learning Models on GPUs — *IPDPS 2020*
Adbul Dakkak, Cheng Li, Jinjun Xiong, Wei Wei, Lingjie Xu, Wen-mei Hwu

The Design and Implementation of the Wolfram Compiler — *CGO 2020*
Adbul Dakkak, Tom Wickham-Jones, Wen-mei Hwu

MLModelScope: A Distributed Platform for Model Evaluation and Benchmarking at Scale — *ArXiv 2020*
Adbul Dakkak, Cheng Li, Jinjun Xiong, Wen-mei Hwu

The Design and Implementation of a Scalable DL Benchmarking Platform — *ArXiv 2019*
Adbul Dakkak, Cheng Li, Jinjun Xiong, Wen-mei Hwu

TrIMS: Transparent and Isolated Model Sharing for Low Latency Deep Learning Inference in Function as a Service Environments — *Cloud 2019*
Adbul Dakkak, Cheng Li, Simon Garcia de Gonzalo, Jinjun Xiong, Wen-mei Hwu

Accelerating Reduction and Scan Using Tensor Core Units — *ICS 2019*
Abdul Dakkak, Cheng Li, Jinjun Xiong, Isaac Gelado, Wen-mei Hwu

Frustrated with Replicating Claims of a Shared Model? A Solution — *ArXiv 2019*
Abdul Dakkak, Cheng Li, Jinjun Xiong, Wen-mei Hwu

Evaluating Characteristics of CUDA Communication Primitives on High-Bandwidth Interconnects — *ICPE 2019*
Carl Pearson, **Abdul Dakkak**, Sarah Hashash, Cheng Li, I Chung, Jinjun Xiong, Wen-mei Hwu

Accelerating Reduction Using Tensor Core Units — *HPCaML 2019*
Adbul Dakkak, Cheng Li, Jinjun Xiong, Wen-mei Hwu

SCOPE: C3SR Systems Characterization and Benchmarking Framework — *ArXiv 2019*
Carl Pearson, **Abdul Dakkak**, Cheng Li, Sarah Hashash, Jinjun Xiong, Wen-mei Hwu

Challenges and Pitfalls of Reproducing Machine Learning Artifacts — *ArXiv 2019*
Cheng Li, **Abdul Dakkak**, Jinjun Xiong, Wen-mei Hwu

TrIMS: Transparent and Isolated Model Sharing for Low Latency Deep Learning Inference in Function as a Service Environments — *SysML NIPS 2018*
Adbul Dakkak, Cheng Li, Simon Garcia de Gonzalo, Jinjun Xiong, Wen-mei Hwu

Thoughts on Massively-Parallel Heterogeneous Computing for Solving Large Problems — *CEM 2017*
Wen-mei Hwu, Mert Hidayetogglu, Weng Cho Chew, Carl Pearson, Simon Garcia, Sitao Huang, **Abdul Dakkak**

RAI: A Scalable Project Submission System for Parallel Programming Courses — *IPDPSW 2017*
Abdul Dakkak, Carl Pearson, Cheng Li, Wen-mei Hwu

WebGPU: A Scalable Online Development Platform for GPU Programming Courses — *IPDPSW 2016*
Abdul Dakkak, Carl Pearson, Wen-mei Hwu

A Programming System for Future Proofing Performance Critical Libraries — *SIGPLAN 2016*
Li-Wen Chang, Izzat El Hajj, Hee-Seok Kim, Juan Gómez-Luna, **Abdul Dakkak**, Wen-Mei W Hwu

Enhancing the Usability and Utilization of Accelerated Architectures via Docker — *UCC 2015*
Nicholas Haydel, Sandra Gesing, Ian Taylor, Gregory Madey, **Abdul Dakkak**, Simon Garcia De Gonzalo, Wen-Mei W Hwu

Massively-Parallel Heterogeneous Computing for Solving Large Problems — *IPDPSW 2016*
Wen-mei Hwu, Mert Hidayetoglu, Carl Pearson, Simon Garcia, Sitao Huang, **Abdul Dakkak**

Tangram: a High-level Language for Performance Portable Code Synthesis — *MULTIPROG*
Li-Wen Chang, **Abdul Dakkak**, Christopher I Rodrigues, Wen-mei Hwu — *2015*

Transitioning HPC software to exascale heterogeneous computing — *CEM*
Wen-Mei Hwu, Li-Wen Chang, Hee-Seok Kim, **Abdul Dakkak**, Izzat El Hajj — *2015*

Triolet: A Programming System that Unifies Algorithmic Skeleton Interfaces for High-performance Cluster Computing — *PPoPP*
Christopher Rodrigues, Thomas Jablin, **Abdul Dakkak**, Wen-mei Hwu — *2014*

Recovering Missing Depth Information from Microsoft's Kinect — *EVA*
Abdul Dakkak, Ammar Husain — *2012*

CUDA & Heterogeneous Programming with the Wolfram Language — *Wolfram Whitepaper*
Abdul Dakkak, Ulises Cervantes-Pimentel — *2012*

CUDA Programming Using Wolfram Finance Platform — *Wolfram Whitepaper*
Abdul Dakkak, Ulises Cervantes-Pimentel — *2011*

Presentations

Using Tensor Cores for Accelerating Reduction and Scan — *San Jose, CA*
GPU Technology Conference — *Mar. 2020*

MLPerf-Bench Tutorial — *Lausanne, Switzerland*
ASPLOS — *Mar. 2020*

Challenges and Solutions for End-to-End and Across Stack ML Benchmarking Tutorial — *Denver, CA*
Super Computing — *Aug. 2019*

Challenges and Solutions for End-to-End and Across Stack ML Benchmarkin Tutorial — *Orlando, FL*
IISWC — *Aug. 2019*

MLModelScope: Evaluate and Profile ML Models at Scale and Across Stack — *Pala Alto, CA*
Hot Chips — *Aug. 2019*

MLPerf-Bench Tutorial — *Pheonix, AZ*
ISCA — *Jun. 2019*

MLPerf-Bench Tutorial — *Providence, RI*
ASPLOS — *Apr. 2019*

MLModelScope — *Champaign, IL*
MLPerf December Meeting 2018 — *Dec. 2018*

Advanced Compilation Techniques — *Champaign, IL*
Wolfram Technology Conference — *Oct. 2018*

TensorOps: Accelerating Reduction Using Tensor Core Units — *San Jose, CA*
NVIDIA GPU Technology Conference — *Mar. 2019*

TrIMS: Transparent and Isolated Model Sharing for Low Latency Deep Learning Inference — *San Jose, CA*
NVIDIA GPU Technology Conference — *Mar. 2019*

MLModelScope: Evaluate and Measure Machine Learning Models within AI Pipelines — *San Jose, CA*
NVIDIA GPU Technology Conference — *Mar. 2019*

MLModelScope: Evaluate and Measure Machine Learning Models within AI Pipelines — *Dallas, TX*
Super Computing — *Nov. 2018*

TRIMS: Transparent and Isolated Model Sharing for Low Latency Deep Learning Inference in Function as a Service Environments — *Boston, MA*
IBM AI Systems Day — *Oct. 2018*

RAI: A Scalable Submission System for GPU Applications — *San Jose, CA*
NVIDIA GPU Technology Conference — *Mar. 2018*

CarML: Common Artifacts for Machine Learning — *Denver, CO*
Super Computing — *Nov. 2017*

Teaching Experience

2018	**Lead TA**, Pumps-AI Summer School	*Barcelona, Spain*
2017	**Lead TA**, CS 508: Manycore Parallel Programming	*Champaign, Illinois*
2016	**Head TA**, Pumps Summer School	*Barcelona, Spain*
2016	**Support TA**, CS 408: Applied Parallel Programming	*Champaign, Illinois*
2015	**Head TA**, Pumps Summer School	*Barcelona, Spain*
2015	**Head TA**, Coursera Heterogeneous Parallel Programming	*Champaign, Illinois*
2014	**Head TA**, Coursera Heterogeneous Parallel Programming	*Champaign, Illinois*
2013	**Head TA**, Coursera Heterogeneous Parallel Programming	*Champaign, Illinois*
2013	**Lead TA**, CS 408: Applied Parallel Programming	*Champaign, Illinois*

Skills

Programming C/C++, Mathematica, CUDA, OpenCL, GoLang, OpenMP, JavaScript, Python, Haskell, LaTeX

Languages English, Arabic