Rafael Andres Herrera Guaitero

randreshg | randres-herrera | Rafael

I am a Ph.D. candidate in Electrical and Computer Engineering at the University of Delaware. I design agentic AI, compiler, and runtime systems that improve performance and reliability in heterogeneous and distributed applications by making execution semantics explicit. My research formalizes execution contracts over State, Dependency, and Effect (S,D,E), then compiles or checks these contracts to automate asynchronous overlap, distributed execution, and multi-step workflow orchestration. In practice, I co-design LLVM/MLIR compiler pipelines with OpenMP/ARTS runtimes and typed agentic workflow IRs to improve latency, utilization, correctness, and developer productivity across HPC and AI systems.

My experience spans industry and national labs. At Meta, I built contract-aware agentic orchestration and privacy-aware data pipelines for internal infrastructure workflows. At Argonne (ECP SOLLVE), I developed task-dependency graph discovery and OpenMP optimization workflows for exascale systems. At PNNL (AMAIS), I prototyped advanced-memory placement workflows for AI-science workloads. At BSC (ODOS), I implemented OpenMP target-offload optimizations for distributed processing units.

Education

2021–Present **Ph.D., Electrical and Computer Engineering**, *University of Delaware*.
Dissertation: Making Execution Semantics Explicit: Automating Distribution and Orchestration Through Execution Contracts.
Advisor: Xiaoming Li **Co-advisers:** Jose M. Monsalve Diaz, Joseph Manzano **Group:** CAPSL

2015–2020 **Bachelor of Engineering**, *Pontificia Universidad Javeriana*, Bogota D.C. - Colombia.
Thesis: "Synthesis of a scene using ray tracing in FPGA"
Advisor: Dr. Juan Carlos Giraldo

Doctoral Research

Title *Making Execution Semantics Explicit: Automating Distribution and Orchestration Through Execution Contracts*

Supervisor Xiaoming Li, Ph.D.

Abstract This dissertation introduces execution contracts, explicit declarations of State, Dependency, and Effect (S,D,E) that make scheduling-relevant semantics compiler-visible across HPC and agentic AI systems. It validates this approach through three pillars: C1 automatic host-device overlap extraction for OpenMP offloading; C2 CARTS, an MLIR-based compiler pipeline that transforms OpenMP into ARTS event-driven execution with 1.42x geometric-mean speedup across 26 benchmarks (up to 5.46x); and C3 A-PXM, a typed contract framework for agentic orchestration with measurable infrastructure gains in latency, code complexity, and contract-error detection.

Bachelor Thesis

Title *Synthesis of a scene using ray tracing in FPGA*

Co-Authors Daniel Saenz, Juan Manuel Gomez.

Advisor Juan Carlos Giraldo, Ph.D.

Co-Advisor Andres Eduardo Maldonado, M.Sc.

Summary Designed and implemented a functional FPGA-based ray tracer and a software reference to analyze bottlenecks. Identified hotspots (intersection tests) and demonstrated up to 18× speedups over the software baseline across test scenes.

Experience

Research and Academic Experience

2022 - Now **Teaching Assistant**, *University of Delaware*, Newark, DE.
Delivered labs, office hours, and grading across core systems and security courses (Computer Networks, Microprocessor Systems, Computer System Design, Cybersecurity, Penetration Testing), improving project readiness and systems-programming fundamentals.

Spring 2026 **Research Associate - PhD**, *AMD Research and Advanced Development*, San Jose, California.
Explored compiler and runtime co-design strategies for heterogeneous execution, focusing on task and memory placement mechanisms that improve efficiency and scalability for emerging AI/HPC workloads.

Summer 2025 **PhD Software Engineer Intern**, *Meta*, Bellevue, WA.
Designed contract-aware agentic orchestration for internal tools and implemented privacy-aware data pipelines for secure infrastructure-policy workflows, improving traceability and reducing manual coordination across services.

Summer 2024 **PhD Intern**, *Pacific Northwest National Laboratory*, Richland, WA.
Developed compiler/runtime prototype workflows for advanced memory platforms in AI-science workloads, translating access-pattern analysis into placement-policy experiments that informed platform design tradeoffs. *Supervisors: Dr. Joseph Manzano, Dr. Andres Marquez*

Spring 2024 **PhD Intern**, *Barcelona Supercomputing Center*, Barcelona, Spain.
Implemented OpenMP target-offload optimizations for distributed processing units, improving offload-path efficiency and overlap opportunities in heterogeneous execution pipelines. *Supervisors: Dr. Antonio Peña, Dr. Sergio Iserte*

Fall 2023 **PhD Intern**, *Pacific Northwest National Laboratory*, Richland, WA.
Built memory-placement analysis tooling and benchmark pipelines for AI-science workloads on advanced memory platforms, enabling reproducible evaluation of tiered-memory behavior. *Supervisors: Dr. Joseph Manzano, Dr. Joshua Sutterlein*

Summer 2023 **PhD Intern**, *Argonne National Laboratory*, Lemont, IL.
Developed OpenMP task-dependency graph discovery analyses and related compile-time/runtime optimizations for exascale systems, enabling earlier identification of schedulable parallelism. *Supervisor: Dr. Jose M. Monsalve Diaz*

Summer 2022 **PhD Intern**, *Argonne National Laboratory*, Lemont, IL.
Implemented automatic asynchronous execution optimizations for OpenMP target offloading in LLVM/OpenMP runtime workflows, contributing to published results with 5–34% speedups (up to 2x in idealized overlap scenarios). *Supervisors: Dr. Jose M. Monsalve Diaz, Dr. Johannes Doerfert*

2021–2022 **Research Assistant**, *CAPSL Research Group*, University of Delaware.
Developed runtime mechanisms and compiler dialect support for the Codelets execution model on embedded multicore systems, enabling explicit dataflow-style decomposition and scheduling experiments. *Supervisors: Dr. Guang R. Gao, Dr. Xiaoming Li, Dr. Jose M. Monsalve Diaz*

Leadership Experience

2024-2025 **President**, *HLGSA & SACNAS*, University of Delaware, Newark, DE.
Lead Hispanic Latino Graduate Student Association and SACNAS initiatives promoting diversity, inclusion, and professional development in STEM.

Awards

2020 **José Gabriel Maldonado Award**, *Pontificia Universidad Javeriana*, Recognition for academic excellence, service, and leadership..

2020 **Outstanding Undergraduate Thesis**, *Pontificia Universidad Javeriana*, Awarded for FPGA-based ray tracing thesis..

Service and Activities

Organizing Committees and Roles

2026 **Tenth LLVM Workshop at CGO**, *Organizing committee*, Collocated with the Code Generation and Optimization (CGO) conference.

2025 **Seventh International Workshop on Emerging Parallel Distributed Runtime Systems and Middleware**, *Publicity Chair*, (IPDRM 2025), collocated with SC24.

2025 **54th International Conference on Parallel Processing**, *Publicity Chair*, (ICPP 2025).

2025 **Ninth LLVM Workshop at CGO**, *Organizing committee*, Collocated with the Code Generation and Optimization (CGO) conference.

2024 **HPC Summer School 2024**, *Organizing committee*, CyberColombia: Building advanced digital infrastructure capacity for research in Colombia.

2024 **Eighth LLVM Workshop at CGO**, *Organizing committee*, Collocated with the Code Generation and Optimization (CGO) conference.

2023 **International Symposium on Cluster, Cloud and Internet Computing**, *Artifact Evaluator*, (CCGRID 2023).

2022 **Americas HPC Collaborations Workshop**, *Artifact Evaluation*, Collocated with the CARLA 2022 conference.

Students

Advisor

(B.Sc.) **Juanita Marulanda Argüello, Juan Pablo Mora Páez, Juan Camilo Bernal Flórez**, *Automatic Fish classification by sound using machine learning techniques*, (Graduated in 2023).

(B.Sc.) **Nicolas Andrey Rios Lopez**, *Scene generation with ray tracing in FPGA*, (Graduated in 2023).

(B.Sc.) **Diego Alejandro Varela Angel**, *GPU ray tracing for 3D model visualization*, (Graduated in 2023).

Skills

Systems and HPC	Agentic AI, LLVM, MLIR, compilers, runtime systems; OpenMP, MPI; dataflow-inspired execution; profiling/perf, sanitizers.	Programming	C (advanced), C++ (intermediate), Python (intermediate), Java (intermediate), Rust (intermediate)
Platforms	Linux, Git, CMake; NUMA and heterogeneous memory; OpenMP target offloading.	CS Foundations	Algorithms, data structures, software design, concurrency, networking basics.
Embedded/HDL	VHDL/Verilog, computer systems design, microcontrollers, electronic circuits.	MCU Platforms	ESP, nRF, PIC, Arduino; sensor integration, comms protocols, real-time control.

Selected Impact

Distributed HPC Compilation	CARTS delivered 1.42x geometric-mean speedup across 26 benchmarks (19/26 wins), with up to 5.46x on a 64-core AMD EPYC system.
Agentic Orchestration	A-PXM achieved up to 10.37x workflow-latency reduction, 7.3x code-complexity reduction, and 49x faster contract-error detection versus LangGraph in controlled-backend evaluations.
OpenMP Offloading	Automatic asynchronous execution from implicit host-device contracts yielded 5–34% speedups on proxy/scientific workloads, with up to 2x in idealized overlap cases.

Publications

Conference Publications

[1] D. A. Roa Perdomo, R. A. Herrera Guaitero, D. Fox, H. Yviquel, S. Raskar, X. Li, and J. M. Monsalve Diaz, "Towards fault tolerance and resilience in the sequential codelet model," in *10th Latin American High Performance Computing Conference (CARLA 2023)*, 2023.

Workshop Publications

[2] R. Herrera Guaitero, J. M. Monsalve Diaz, T. Applencourt, J. Doerfert, and X. Li, "Automatic asynchronous execution of synchronously offloaded openmp target regions," in *The Eighth Annual Workshop on the LLVM Compiler Infrastructure in HPC, 2022 (LLVM-HPC)*, 2022.

[3] J. M. M. Diaz, K. Harms, R. A. H. Guaitero, D. A. R. Perdomo, K. Kumaran, and G. R. Gao, "The supercodelet architecture," in *Proceedings of the 1st International Workshop on Extreme Heterogeneity Solutions*, ser. ExHET '22, 2022. [Online]. Available: https://doi.org/10.1145/3529336.3530823

Thesis

[4] G. Herrera, R. A., J. M., and D. A. Sáenz, "Synthesis of a scene using ray tracing on fpga," 2019. [Online]. Available: http://hdl.handle.net/10554/57527

Selected Work in Progress

Status snapshot: February 13, 2026.

P5 **CARTS: Distributed Contract Extraction from OpenMP into Event-Driven ARTS Execution**, *Euro-Par 2026*, Status: Active submission phase (First Author).
MLIR-based compilation pipeline for extracting State, Dependency, and Effect contracts from OpenMP.

P4 **A-PXM: Contract-Typed Agentic Orchestration with Dataflow Scheduling**, *Computer Frontiers 2026*, Status: Under review (First Author).
Typed state, dependency, and effect contracts for agent workflows with compile-time checking.

P0 **Separation Principle for Tasking Models**, *Unpublished Position Draft*, Status: Draft formalization (First Author).
Articulates control-flow and task-graph concern conflation motivating explicit contract IRs.

P6 **CARTS Extension for Disaggregated Memory Contracts**, *Target venue: TBD*, Status: Planned post-defense (First Author).
Planned extension of contract extraction to placement-affinity and memory-tier boundaries.

P7 **CARTS Extension for GPU Execution Contracts**, *Target venue: TBD*, Status: Planned post-defense (First Author).
Planned extension unifying distributed and GPU execution under the same contract substrate.

Talks and Presentations

CGO26 **Compiling Agentic AI Programs for Dataflow Execution with Explicit Contracts**, *10th LLVM Performance Workshop at CGO*, Las Vegas, USA, February 2026.

CGO25 **CARTS: Enabling Event-Driven Task and Data Block Compilation for Distributed HPC**, *9th LLVM Performance Workshop at CGO*, Las Vegas, USA, March 2025.

CGO24 **Unveiling the Power of Heterogeneous Computing: A Brief Dive into Host and Target Tasks with OpenMP LLVM**, *8th LLVM Performance Workshop at CGO*, Edinburgh, United Kingdom, March 2024.

WAMTA2024 **On Scheduling IRs for Program Execution Models**, *Workshop on Asynchronous Many-Task Systems and Applications*, Knoxville, TN, USA, February 2024.

LLVM23 **TDG discovery and compile-time optimizations of OpenMP Tasks**, *2023 LLVM Developers' Meeting*, Santa Clara, California, USA, October 2023.

LLPP22 **Automatic Asynchronous Execution of Synchronously Offloaded OpenMP Target Regions**, *The Second Workshop on LLVM in Parallel Processing*, Virtual, May 2022.