

# Quinn Leo Pham

he/him  
+1 (587) 573-7731  
Edmonton, AB  
[qpham@ualberta.ca](mailto:qpham@ualberta.ca)  
[github.com/quinnlp](https://github.com/quinnlp)  
[linkedin.com/in/quinnlp](https://linkedin.com/in/quinnlp)

## Education

---

Sep 2025 – Present **P.h.D. Computing Science**

University of Alberta, Edmonton, Canada

- Supervisor: José Nelson Amaral

Sep 2023 – Sep 2025 **M.Sc. Thesis Computing Science**

University of Alberta, Edmonton, Canada

- Thesis: *Decoupled Triton: A Block-Level Decoupled Language for Writing and Exploring Efficient Machine-Learning Kernels*
- GPA: 4.0 / 4.0
- Supervisor: José Nelson Amaral

Sep 2018 – Apr 2023 **B.Sc. Honors Computing Science, Science Internship Program**

*with Gold Medal in Computing Science, Dean's Silver Medal in Science, and First Class Honors*

University of Alberta, Edmonton, Canada

- GPA: 3.8 / 4.0

## Research Experience

---

May 2023 – Present **Graduate Researcher**

University of Alberta, Edmonton, Canada

- Decoupled Triton: a block-level decoupled domain specific language for writing machine-learning kernels that enables rapid schedule exploration.
- Routed and Cascaded Deep Neural Network: a routed dynamic deep neural network architecture that reduces inference time over an early-exit cascaded deep neural network by reducing wasted computation.
- Tensor Shape-Specialized Adaptive Cache: a kernel compilation and caching system that improves performance by adaptively compiling shape-specialized or shape-generic kernels.
- Code Generation for Branch Prediction Literature Review: a literature review of code generation strategies for improving branch prediction accuracy.

May 2021 – Apr 2023 **Compiler Researcher**

IBM, Toronto, Canada

- Dynamic Adaptive Sub-Target Specialization: a compiler system that utilizes a fat-binary and dynamic compilation to access optimization opportunities available on new architecture versions that are inaccessible by a generic static compilation.

May 2020 – Aug 2020 **Undergraduate Researcher**

University of Alberta, Edmonton, Canada

- Active Lane Consolidation: Developed a compiler pass to identify loops with divergent control flow as targets for a vector instruction optimization that consolidates the active vector lanes of multiple vector registers into a single vector register.

## Teaching Experience

---

Sep 2023 – Dec 2023 **Compiler Design (CMPUT 415)**

*Teaching Assistant*

University of Alberta, Edmonton, Canada

- Received 100% positive feedback from students about the instructional approach and class climate, guided student teams in completing a front-end compiler for a C-like language targeting LLVM Dialect MLIR, prepared and presented tutorials and presentations, monitored and resolved intra-team conflicts, answered questions on the course forums, and provided feedback to and evaluated students.

Sep 2020 – Dec 2020 **Computer Organization and Architecture I (CMPUT 229)**

*Teaching Assistant*

University of Alberta, Edmonton, Canada

- Led help sessions, answered questions on the course forums, evaluated and provided feedback to students, and investigated possible academic integrity violations.

## Awards

---

Sep 2025 – Aug 2026 **Graduate Recruitment Scholarship**, \$5,000

Sep 2024 – Aug 2025 **Science Graduate Scholarship**, \$2,000

Sep 2023 – Aug 2024 **NSERC Canada Graduate Scholarship - Master's Program**, \$17,500

Sep 2023 – Aug 2024 **Walter H Johns Graduate Fellowship**, \$7,100

Sep 2023 – Aug 2024 **Graduate Recruitment Scholarship**, \$5,000

Apr 2023 **Gold Medal in Computing Science**

*Highest average in courses taken in the last three years of the program.*

Apr 2023 **Dean's Silver Medal in Science**

*Superior academic achievement.*

Apr 2023 **First Class Honors**

Nov 2022 **LLVM Foundation Travel Grant**

Nov 2021 **Jason Lang Scholarship**, \$1,000

Sep 2020 **Louise McKinney Post-Secondary Scholarship**, \$2,500

*Superior academic achievement (top 1.5-2% of faculty).*

Sep 2020 **University of Alberta Undergraduate Scholarship**, \$2,000

*Superior academic achievement.*

Sep 2020 **CIPS Stan Heaps Memorial Scholarship**, \$2,500

*Superior academic achievement.*

May 2020 – Aug 2020 **NSERC Undergraduate Student Research Award**, \$4,500

Jul 2020 **Jason Lang Scholarship**, \$1,000

Aug 2018 **Alexander Rutherford High School Achievement Scholarship**, \$2,500

## Publications

---

### Peer-Reviewed

1. T. Gobran, **Q. Pham**, J. P. L. de Carvalho, J. N. Amaral, C. Barton, and N. Ivanovic, "DASS: Dynamic Adaptive Sub-Target Specialization," *2023 International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW)*, Porto Alegre, Brazil, 2023, pp. 36-45, doi: 10.1109/SBAC-PADW60351.2023.00016.

## Conference Abstracts

---

### Posters

1. **Q. Pham**, D. Seliayeu, P. Chatarasi, and J. N. Amaral, “Decoupled Triton: Exploring Coupled and Decoupled Machine-Learning Kernel Languages,” *2024 Collaborative Advances in Software and Computing (CASCON)*, Toronto, Canada, 2024, link: [webdocs.cs.ualberta.ca/qpham/resources/posters/DT-CASCON24.pdf](http://webdocs.cs.ualberta.ca/qpham/resources/posters/DT-CASCON24.pdf).
2. D. Seliayeu, **Q. Pham**, P. Chatarasi, and J. N. Amaral, “Mixture of Shared Experts,” *2024 Collaborative Advances in Software and Computing (CASCON)*, Toronto, Canada, 2024.
3. D. S. Hira, **Q. Pham**, T. Gobran, J. P. L. de Carvalho, N. Ivanovic, C. Barton, and J. N. Amaral, “Specializing Code to New Architectures via Dynamic Adaptive Recompilation,” *2022 LLVM Developers’ Meeting (LLVMDev)*, San Jose, USA, 2022, link: [webdocs.cs.ualberta.ca/qpham/resources/posters/DAR-LLVMDev22.pdf](http://webdocs.cs.ualberta.ca/qpham/resources/posters/DAR-LLVMDev22.pdf).

### Presentations

---

#### Seminars

1. **Q. Pham**, “Decoupled Triton: A Block-Level Decoupled Language for Writing and Exploring Efficient Machine-Learning Kernels”, Edmonton, Canada, 2025, link: [webdocs.cs.ualberta.ca/qpham/resources/decks/DT-MSD.pdf](http://webdocs.cs.ualberta.ca/qpham/resources/decks/DT-MSD.pdf).
2. **Q. Pham**, “The Argument for Decoupled Triton”, Edmonton, Canada, 2025, link: [webdocs.cs.ualberta.ca/qpham/resources/decks/TAFDT-MAR25.pdf](http://webdocs.cs.ualberta.ca/qpham/resources/decks/TAFDT-MAR25.pdf).

#### Lectures

1. **Q. Pham**, “Code Generation for Branch Prediction: a review”, *Advanced Compiler Design (CMPUT 680)*, Edmonton, Canada, 2023, link: [webdocs.cs.ualberta.ca/qpham/resources/decks/CGFBP-CMPUT680.pdf](http://webdocs.cs.ualberta.ca/qpham/resources/decks/CGFBP-CMPUT680.pdf).
2. **Q. Pham**, “Neural Branch Predictors”, *Computer Systems and Architecture (CMPUT 529)*, Edmonton, Canada, 2023, link: [webdocs.cs.ualberta.ca/qpham/resources/decks/NBP-CMPUT529.pdf](http://webdocs.cs.ualberta.ca/qpham/resources/decks/NBP-CMPUT529.pdf).
3. **Q. Pham**, “Using LLVM and MLIR”, *Compiler Design (CMPUT 415)*, Edmonton, Canada, 2023, link: [webdocs.cs.ualberta.ca/qpham/resources/decks/ULAM-CMPUT415.pdf](http://webdocs.cs.ualberta.ca/qpham/resources/decks/ULAM-CMPUT415.pdf).
4. **Q. Pham**, “What is LLVM (and MLIR)?”, *Compiler Design (CMPUT 415)*, Edmonton, Canada, 2023, link: [webdocs.cs.ualberta.ca/qpham/resources/decks/WILAM-CMPUT415.pdf](http://webdocs.cs.ualberta.ca/qpham/resources/decks/WILAM-CMPUT415.pdf).

### Service

---

#### Reviewing

- Oct 2024 **Parallel Architectures and Compilation Techniques (PACT)**  
Long Beach, USA
- Jun 2024 **International Symposium on Computer Architecture (ISCA)**  
Buenos Aires, Argentina

#### Volunteering

- Nov 2024 **Collaborative Advances in Software and Computing (CASCON)**  
Toronto, Canada
- Nov 2022 **Conference of the Center for Advanced Studies on Collaborative Research (CASCON)**  
Toronto, Canada

### Professional Development

---

- Sep 2023 – Dec 2023 **Teaching and Research Methods**  
University of Alberta, Edmonton, Canada

## Additional Information

---

Research Interests	Compiler Design, Compiler Optimization, Decoupled Languages, User-Schedulable Languages, Hardware Accelerators, Deep Learning Compilers, Dynamic Deep Neural Networks, Code Generation for Branch Prediction, Neural Branch Predictors, Dynamic Compilation, Binary Optimization, Computer Architecture
Technical Skills	C++, Python, Bash, Triton, PyTorch, CUDA, LLVM, MLIR, JavaScript, C, Halide, Git, UNIX, CMake, ANTLR4, OpenMP, WebGL
Languages	Fluent: English, French Novice: Japanese
Nationality	Canadian Citizen