Brian Won (Injong)

Software Engineer — Systems Programming — ML & Distributed Systems

injongwbrian@gmail.com https://injongwon.github.io/

Education

University of Toronto Toronto, ON

B.Sc. in Computer Science — Specialization in ML & Systems *Class of 2025*

- Teaching Assistant: Operating Systems, Computer Networks, Database Management Systems

Experience

University of Toronto Engineering Toronto, ON

Research Assistant *Apr 2025 – Present*

- Engineered full-stack academic contest platform serving 1,200+ students across Ontario using **React, Express.js** with real-time leaderboard synchronization
- Architected scalable database using **Prisma ORM** with normalized schema, reducing runtime errors by 70% through type-safe operations
- Implemented secure authentication with **bcrypt** and RBAC for Ontario Association of Physics Teachers, achieving 99.5% uptime

Vector Institute Toronto, ON

Research Assistant *Jul 2024 – Dec 2024*

- Developed hybrid operating system in **C/Rust** with custom packet parsers, ARP logic, and TCP-inspired socket protocol
- Implemented cooperative and preemptive scheduling algorithms, reducing context switch latency by 25%

RBC Capital Markets Montreal, QC

Software Engineer (Co-op) *Sep 2023 – Dec 2023*

- Architected distributed ETL pipelines processing 1M+ daily financial transactions, achieving 2x throughput improvement
- Developed real-time monitoring dashboard maintaining 99.9% uptime for $500M+ daily trading volume
- Implemented automated alerting system reducing incident response time by 40%

Analytic Partners New York, NY

Software Engineer (Co-op) *May 2023 – Aug 2023*

- Built **Pandas-based** microservices automating spreadsheet operations, reducing manual processing time by 20%
- Integrated end-to-end testing workflows with QA teams ensuring reliability of revenue-impacting data pipelines

IBM Canada Toronto, ON

Software Engineer (Co-op) *May 2019 – Aug 2020*

- Migrated enterprise **Angular** frontend (15K+ LOC) from v6 to v8, achieving 25% performance improvement
- Implemented automated **CI/CD pipelines** reducing release cycle time by 30% and post-deployment bugs by 35%

Selected Projects

Multi-Head Latent Attention (MLA) Implementation 2025

- Implemented DeepSeek's Multi-Head Latent Attention in **PyTorch**, reducing KV cache memory footprint by 60% while maintaining model performance
- Optimized attention computation with efficient matrix operations, achieving 2.5x speedup on inference benchmarks

High-Performance Operating System Kernel 2024

- Developed custom OS kernel in **C/Rust** with advanced scheduling algorithms and inter-process communication mechanisms
- Achieved 40% reduction in context switching overhead through optimized scheduler design and cache-friendly data structures

Technical Skills

Languages: Python, C/C++, Java, JavaScript, Swift, Ruby, Rust, SQL
Frameworks: React/React Native, Angular, Express.js, Ruby on Rails, TensorFlow, PyTorch
Databases: PostgreSQL, MySQL, MongoDB, Redis, Prisma ORM
Tools: Git, Docker, CI/CD (Jenkins, GitHub Actions), Linux, REST APIs, WebSockets
Systems: Distributed Systems, Network Protocols (TCP/IP, BGP, OSPF), Performance Optimization