Seung Ho, Jang

South Korea | (+82) 10-3550-5156 | shjang1992@hotmail.com | https://github.com/skyser2003

Technical Skills

Language: Python, TypeScript, Go, Rust, C++
LLM inference optimization: vLLM, TensorRT-LLM
Orchestration / Workflow : Kubernetes, Docker, GCP

Work Experience

Software Engineer, MLOps / AI Engineer NCSOFT – South Korea September 2017 – May 2024

- **LLM Optimization and Deployment** (June 2023 – May 2024)
  - Optimized inference latency/throughput for transformer models using TensorRT-LLM/vLLM, improving serving performance in production environments
  - Applied general optimization technics such as model quantization and prefix caching
  - Conducted stress testing and fine-tuning of server configurations, enhancing throughput and system reliability
- **Microservices Migration and Helm Integration** (June 2022 – December 2022)
  - Migrated existing Kustomization-based Kubernetes deployments to Helm, consolidating multi-environment configurations into modular templates
  - Automated CI/CD workflows using ArgoCD, reducing deployment complexity and manual intervention across staging and production clusters
- **Machine Learning Model (CPU) Serving and Optimization** (June 2021 – May 2024)
  - Deployed CPU inferencing ML models using Tritonserver to serve batch and streaming inference requests for high-demand applications
  - Engineered scalable serving infrastructure, addressing latency and stability issues by developing a custom batch inference server in Rust/Go to replace problematic Tritonserver, achieving significant performance gains
- **Machine Learning Platform Development** (June 2019 – May 2021)
  - Designed and implemented a Kaggle-like ML training and evaluation platform with Jupyter Notebook integration, supporting GPU resource allocation and automated leaderboard generation
  - Built a comprehensive ML job management platform with real-time monitoring, log aggregation, and hyperparameter tuning features, leveraging Kubernetes and Elasticsearch
- **GraphDB Application Development** (March 2019 – May 2019)
  - Created a GraphDB application to dynamically update real-time baseball game statistics, utilizing Neo4j for efficient data representation and query optimization
  - Improved database performance by restructuring node and relationship queries, achieving faster updates and searches
- **Web Server Development** (September 2017 – August 2018)
  - Developed backend services for the AI-powered baseball application Paige using Node.js and TypeScript
  - Implemented FCM-based notification services
  - Automated CI/CD pipelines with Jenkins and Docker

Game Developer IMC Games – South Korea September 2012 – August 2015

- **Spirit Wish (Mobile MMORPG)** (March 2015 – August 2015)
  - Developed the server architecture, including world and channel systems, using C# and Protobuf
  - Designed and conducted stress tests with custom dummy clients to ensure server robustness and scalability
- **Granado Espada (PC MMORPG)** (October 2012 – February 2015)
  - Enhanced XML parsing efficiency using multithreaded programming, reducing loading times significantly
  - Built a real-time XML reload system and an automated packet management framework using C++ templates
  - Integrated centralized logging with ELK for centralized monitoring

Education

Yonsei University – B.S. in Computer Science and Engineering, Minor in Psychology March 2010 – February 2017

Language Skills

- **Korean:** Native proficiency
- **English:** Fluent (TOEFL iBT 111, acquired March 2009)
- **French:** Intermediate (DELF B2, acquired June 2007)

Etc. Personal Projects

- **Template Library for C++** (GitHub: FTL): Implemented C#-style properties, Unity-style Vector class, and Enum enumerator in a C++ template library.
- **Personal Deep Learning Project** (GitHub: ML): Used InfoGAN, DCGAN, and Improved WGAN to generate pixel-style game character images.
- **Nexon Developer Conference 2014 Presentation** (NDC 2014): Delivered a presentation on creating projects that are easy to maintain for entry-level programmers.
- **Hololens Demo Project** (GitHub: Artan): Developed a piano lesson program and a 3D artillery game playable on Hololens.
- **Personal Slack Bot** (GitHub: Ditto Bot Rust): Built a Slack bot for web page parsing and preview generation, chat count tracking per user, and ChatGPT API integration using Rust.
- **Monster Hunter Skill Calculator** (GitHub: MHR Skill Calculator): A tool developed with Vue3 and Rust to calculate optimal armor and skill combinations for the game 'Monster Hunter: Sunbreak'.