

Job Title: AI Compiler Engineer (Automotive)

Level: Team Leader/ Fresher/Junior

Company Overview:

BOS Semiconductors is a fabless startup based in South Korea, established in 2022. The company specializes in **designing and developing system-on-chip (SoC) solutions**. Their innovations include high-performance, low-power semiconductors, CPUs, graphics cards, and high-speed signal interfaces, all of which are essential for **automotive** semiconductors. Hyundai Motor Group has invested in BOS Semiconductors, anticipating collaboration towards the semiconductor technologies necessary for future mobility.

BHRC, established in Ho Chi Minh City in November 2022, is a key R&D center that comprises 40% of BOS's total R&D workforce. Over the past two years, the team has primarily focused on **hardware development**, hiring excellent graduates from your university and growing together. BHRC is rapidly expanding its scope and is now extending into **software development**, preparing for the fab-out of its first production SoC. The main areas of work include **developing embedded software for ADAS chips, AI accelerator chips, and MCUs** responsible for functional safety in autonomous vehicles. One significant opportunity for the BHRC software team is to learn about the world's most advanced NPU architecture and design the BUDA and metal-BUDA compilers for this NPU engine.

The NPU vendor collaborating with BHRC is Tenstorrent, led by the renowned Jim Keller. He is famous for creating high-performance CPU architectures at Apple, Intel, AMD, and Tesla and now leads Tenstorrent, a startup developing RISC-V-based NPU engines and Transformer-based server chips. BOS's executives, former SoC developers at Samsung, have a special relationship with Jim Keller, having collaborated on several projects, including the ADAS chip for the Tesla Model 3. This relationship has allowed BOS to incorporate the world's **most advanced A/I engine** as a core IP in its first production chip and to take on the mission of developing the compiler for the most advanced NPU architecture.

BHRC plans to hire and rapidly develop talented graduates from your university, led by a few greatly talented leaders.

Position Overview:

We are seeking a highly skilled and motivated AI Compiler Engineer specializing in NPUs to join our dynamic team. The ideal candidate will have a strong background in compiler theory and design, as well as extensive experience in optimizing AI models for NPU architectures. The AI Compiler Engineer will play a critical role in developing and optimizing AI models for deployment on NPUs, particularly ResNet50 or VGG19, Inception3, EfficientNet, to enhance the performance and efficiency of our automotive systems. Key Responsibilities:

If you are a highly motivated individual who is passionate about designing and developing new technology and wants to make a significant impact in the automotive industry, we encourage you to apply for this exciting opportunity (Even if you do not have any experience below) list below:

1. Develop and optimize compiler tools and runtime systems specialized for NPUs, focusing on ResNet50 and other deep learning models.
2. Collaborate with the AI research and development team to understand the architecture of the NPU and develop strategies for optimizing AI models for this hardware.

3. Design, implement, and optimize compiler algorithms and runtime libraries to enable efficient execution of AI models on NPUs.
4. Work closely with the AI engineering team to integrate AI models into our automotive systems and ensure optimal performance on NPU hardware.
5. Analyze and evaluate the performance of AI models on NPUs, identify bottlenecks, and propose solutions for optimization.
6. Develop and maintain documentation for compiler tools, runtime systems, and optimization techniques.
7. Stay up to date with the latest advancements in compiler technology and AI hardware architecture to continuously improve the performance and efficiency of our AI models.

Qualifications:

- Bachelor's, Master's, or Ph.D. degree in Computer Science, Electrical Engineering, or related field. (GPA at least 8.0).
- Proven experience in AI/ML development, with a focus on neural network design, optimization, and deployment (Good to have).
- In-depth understanding of NPU architectures, including familiarity with hardware/software co-design principles (Good to have).
- Proficiency in programming languages commonly used in AI development (e.g., Python, TensorFlow, **PyTorch**) (Good to have).
- Proficiency in programming languages such as C, C++, CUDA C++, and Python or assembly language.
- Strong mathematical background, including knowledge of linear algebra and calculus.
- Experience with deep learning frameworks such as TensorFlow, PyTorch, or similar.
- Familiarity with ResNet50 and other popular deep learning models.
- Excellent problem-solving skills and the ability to analyze and debug complex software systems.
- Strong communication and collaboration skills, with the ability to work effectively in a dynamic, cross-functional team environment.
- Experience in the automotive industry is a plus.

How to Apply:

Interested candidates should submit their resume and a cover letter detailing their relevant experience and qualifications to trangpham@bos-semi.com

Please include "AI Engineer (NPU) Application" in the subject line. We look forward to reviewing your application and discussing how you can contribute to our innovative team.