【日系半導体】Principal AIコンパイラエンジニア｜3nm SoC・NPU

英語中心のグローバルな環境で、専門性とリーダーシップを磨きたい方に

Job Information

Recruiter
Morgan McKinley

Job ID
1619177

Industry
Bank, Trust Bank

Job Type
Permanent Full-time

Location
Tokyo - 23 Wards

Salary
Negotiable, based on experience

Refreshed
November 9th, 2026 00:00

General Requirements

Minimum Experience Level
Over 3 years

Career Level
Mid Career

Minimum English Level
Fluent Level

Minimum Japanese Level
Basic Level

Minimum Education Level
Bachelor's Degree

Visa Status
Permission to work in Japan required

Job Description

自動車・産業機器向けの先端半導体を手がける日系グローバル企業で、次世代3nm SoC向けAIコンパイラの設計・開発・性能最適化をリードするポジションです。

AIコンパイラのアーキテクチャ設計からNPUの性能チューニング、ハードウェアとの協調設計まで幅広い技術領域で裁量を発揮し、自動運転を支える次世代車載AIの実現に貢献できます。

英語中心のグローバルな環境で、専門性とリーダーシップを磨きたい方におすすめのポジションです。

主な職務内容

- モデルの取り込み、計算グラフの最適化、低レベル表現への変換（Lowering）、コード生成、ランタイム統合に至るAIコンパイラのアーキテクチャ全体を主導する。
- 演算子融合（Operator Fusion）、量子化を考慮したグラフ変換、メモリ効率を考慮したスケジューリング、処理の

分割など、計算グラフレベルの最適化技術を設計・実装する。
- タイリング、テンソルレイアウトの最適化、マルチコア実行戦略を通じて、NPUの性能向上を推進する。
- SoCおよびAIアクセラレータのアーキテクトと連携し、コンパイラの分析・知見をハードウェア機能の設計に反映する。
- シミュレータ、プロファイラ、実シリコンとの相関を考慮したモデルを活用し、車載AIワークロードの性能指標（KPI）を管理・改善する。
- 機能安全、リアルタイム性、決定論性に関する要件を満たすコンパイラ出力を実現する。
- シニアエンジニアの技術的成長を支援し、直接的な人員管理を伴わずに技術方針の策定と専門領域のリーダーシップを担う。

Join a **global semiconductor technology leader** as a **Principal AI Compiler Engineer**, driving the architecture and development of advanced compiler and code generation technologies for next-generation **3nm System-on-Chip (SoC) platforms**.

In this permanent position, you will play a central role in maximising compute performance for **automotive AI, neural processing units (NPUs) and hardware accelerators**, supporting the development of software-defined vehicles and next-generation smart mobility technologies.

You will lead the design of **AI compiler architectures**, covering model ingestion, computational graph optimisation, intermediate representation lowering, code generation and runtime integration.

Working at the intersection of **AI compiler engineering, machine learning inference and hardware-software co-design**, you will optimise the execution of complex AI workloads on highly specialised semiconductor architectures.

Your work will include **operator fusion, quantisation-aware graph transformations, memory-aware scheduling, tensor layout optimisation, tiling and multi-core execution strategies**, with a strong focus on performance, efficiency and deterministic execution.

You will collaborate with **SoC architects, AI accelerator designers and high-performance computing engineers** to influence future hardware capabilities and ensure compiler output meets demanding automotive requirements for **functional safety, real-time performance and reliability**.

As a principal-level technical specialist, you will define engineering direction, mentor experienced engineers and establish best practices without direct people management responsibilities.

This opportunity is ideal for an experienced **AI Compiler Engineer, Compiler Architect, NPU Compiler Engineer, Machine Learning Compiler Specialist or Semiconductor Software Engineer** seeking a high-impact permanent position in Tokyo.

Key Responsibilities

- Lead the architecture and development of **AI compiler technology** across model ingestion, graph optimisation, intermediate representation lowering, code generation and runtime integration.
- Design and implement advanced **computational graph optimisations**, including operator fusion, quantisation-aware rewrites, memory-aware scheduling and graph partitioning.
- Optimise **NPU and AI accelerator performance** through tiling, tensor layout optimisation, memory management and multi-core execution strategies.
- Develop and improve compiler backends and code generation techniques for **custom hardware accelerators and next-generation 3nm SoC platforms**.
- Collaborate with **SoC and AI accelerator architects** to influence hardware design decisions using insights from compiler behaviour and workload performance.
- Define, measure and improve **performance KPIs for automotive AI inference workloads**, using simulators, profilers and silicon-correlated performance models.
- Ensure compiler-generated code meets **functional safety, deterministic execution and real-time processing requirements** for automotive applications.
- Evaluate and enhance compiler infrastructure using technologies such as **MLIR, LLVM, TVM and proprietary intermediate representations (IRs)**.
- Work with cross-functional hardware and software engineering teams to optimise **CNNs, transformers, perception models and generative AI inference workloads**.
- Provide **principal-level technical leadership and mentorship** to senior engineers, establishing architectural direction and engineering best practices without direct line management responsibilities.

Required Skills

必須条件 経験・資格：

- コンピュータサイエンス、電気電子工学、または関連分野の修士号・博士号、もしくはそれに相当する実務経験。
- AIコンパイラ、アクセラレータ向けバックエンド、または計算グラフ最適化フレームワークの開発経験。
- NPUまたは独自のハードウェアアクセラレータを対象とした、計算グラフの最適化および性能チューニングの実務経験。
- MLIR、LLVM、TVM、または独自のコンパイラ中間表現（IR）を用いた開発経験。
- CNN、Transformer、認識系AI、生成AIモデルなど、AI推論ワークロードに関する確かな知識。
- C/C++およびPythonの高度なプログラミングスキル。

ソフトスキル：
- Scrumなどのアジャイル開発環境における、優れたコミュニケーション能力。

- ビジョンを策定し、シニアエンジニアの育成・指導を通じてチーム全体の技術力向上を推進できるリーダーシップ。
- **エンジニアリング品質、性能最適化、技術革新**に主体的に取り組むオーナーシップ。

語学力：
- **日本語：**できれば尚可
- **英語：**流暢

歓迎条件
- 自動車向けソフトウェア、またはセーフティクリティカルなシステムの開発経験。
- CPU、GPU、DSP、NPUなどを組み合わせたヘテロジニアスSoCアーキテクチャに関する知識・経験。
- 性能モデリング、またはハードウェアとソフトウェアの協調設計（Hardware-Software Co-design）の経験。

採用企業について
日系半導体メーカー。

自動車、産業機器、社会インフラ向けに、高度なSoCおよびマイクロコントローラソリューションを開発するグローバル企業です。

同社のモビリティ部門は、ソフトウェア定義型車両（SDV）や次世代スマートモビリティを支える、最先端の3nmシリコンプラットフォームの開発を推進しています。

世界規模の技術開発体制のもと、先進的な半導体アーキテクチャとAI技術を融合させ、次世代自動車の高度化を支えています。

この求人がおすすめの理由
- **次世代の車載AI技術を支える：** 最先端の3nm SoC向けコンパイラ技術を主導し、自動運転や高度なAIワークロードの実現に直接貢献できます。
- **裁量の大きなポジション：** AIコンパイラのアーキテクチャ設計から性能最適化、ハードウェアとの協調設計まで、幅広い領域で技術的な意思決定をリードできます。
- **世界水準のHPCエンジニアリング組織：** 高性能コンピューティングと先端半導体技術の専門家と連携し、次世代の車載AIを支える技術課題に取り組めます。
- **技術リーダーシップを発揮：** シニアエンジニアへの技術指導や技術方針の策定を通じて、チーム全体の技術力向上に貢献できます。
- **英語中心のグローバルな環境：** 英語力を活かし、国際的なビジネススキルを磨けます。

Required Skills and Qualifications**Experience:**
- **Master's degree or PhD in Computer Science, Electrical Engineering or a related discipline** , or equivalent professional experience.
- Deep professional experience designing and developing **AI compilers, machine learning compiler frameworks, accelerator backends or graph optimisation systems**.
- Strong hands-on expertise in **computational graph optimisation and performance tuning for NPUs or custom AI hardware accelerators**.
- Practical experience with **MLIR, LLVM, TVM or proprietary compiler intermediate representations (IRs)** .
- Strong understanding of **AI inference workloads**, including convolutional neural networks (CNNs), transformers, perception models or generative AI.
- Advanced programming proficiency in **C, C++ and Python** for compiler development, performance optimisation and systems engineering.
- Solid understanding of **compiler code generation, model lowering, runtime integration and hardware-aware optimisation**.
- Experience analysing performance bottlenecks and improving execution efficiency on specialised compute architectures.

Soft Skills:
- Strong **technical leadership and architectural vision** , with the ability to define engineering direction for complex compiler technologies.
- Excellent communication skills when collaborating with **semiconductor architects, AI researchers and software engineering teams**.
- Proven ability to **mentor senior engineers and influence technical decisions** without direct people management authority.
- Analytical and methodical problem-solving skills when investigating complex compiler and hardware performance challenges.
- Ability to work effectively within **Agile development environments**, including Scrum-based engineering teams.
- Strong ownership mindset with a commitment to **engineering quality, performance optimisation and technical innovation**.

Language Requirements:
- **Japanese:** Basic level for communication within the local working environment.
- **English:** Fluent level for technical discussions, engineering documentation and collaboration with global development teams.

Preferred Skills & Qualifications
- Experience developing **automotive software, autonomous driving systems or other safety-critical applications** .
- Strong understanding of **heterogeneous SoC architectures**, including CPU, GPU, DSP and NPU processing units.
- Experience in **hardware-software co-design**, particularly where compiler requirements influence accelerator architecture.
- Knowledge of **performance modelling, simulator-based profiling and silicon-correlated performance analysis** .
- Experience optimising **AI inference workloads for embedded systems, real-time processing or automotive perception applications**.
- Familiarity with **functional safety, deterministic execution and performance constraints** in automotive semiconductor environments.

About the Company
Our client is a **global leader in semiconductor technology**, developing advanced **System-on-Chip (SoC) and**

microcontroller solutions for automotive, industrial and infrastructure applications.

The organisation operates at the forefront of semiconductor innovation, combining sophisticated silicon design with advanced software engineering to support increasingly demanding computing requirements.

Its mobility division is pioneering **next-generation 3nm silicon platforms** designed to power **software-defined vehicles, advanced automotive AI and smart mobility technologies**.

The company's engineering teams work across **high-performance computing, AI acceleration, compiler development and hardware-software integration**, enabling complex machine learning workloads to run efficiently on specialised semiconductor platforms.

With a strong focus on technical excellence and long-term innovation, the organisation provides an environment where experienced engineers can shape foundational technologies and influence future generations of automotive computing systems.

This is an opportunity to lead **core AI compiler technology for flagship semiconductor platforms**, contributing directly to the evolution of autonomous driving and intelligent mobility.

Why You'll Love Working Here

- Take a leading role in developing **core AI compiler technology for flagship 3nm SoC platforms**.
- Contribute directly to **next-generation autonomous driving, automotive AI and software-defined vehicle technologies**.
- Work on advanced **NPU optimisation, compiler code generation and hardware accelerator performance engineering**.
- Influence the design of future semiconductor hardware through **compiler-driven hardware-software co-design**.
- Gain exposure to sophisticated **MLIR, LLVM, TVM and AI compiler optimisation techniques**.
- Solve challenging problems involving **real-time AI inference, functional safety and deterministic execution**.
- Collaborate with a **world-class high-performance computing and semiconductor engineering team**.
- Enjoy substantial **technical influence and architectural ownership** without requiring direct people management responsibilities.
- Mentor experienced engineers and help define **compiler architecture standards and engineering best practices**.
- Build a long-term career at the forefront of **AI semiconductor innovation, 3nm SoC development and next-generation automotive computing**.

Company Description

Morgan McKinleyは国際的な人材コンサルティング会社として、さまざまな業界・分野をリードする採用企業様と、スペシャリストとしてのスキルを有する人材とを結びつけるお手伝いをしています。1988年の創立以来、Morgan McKinleyの名は、「卓越した質のサービス」「市場知識の豊富さ」「No.1企業であり続けようとする強い意志」、そして何よりも「実績」を体現する会社として知られています。

リクルーティング スペシャリストである弊社コンサルタントまでお気軽にお問い合わせください。