PROFILE

Director of Architecture
Ambarella

CITATION

Technical leader advancing artificial intelligence from silicon architecture to systems

MAX FANG

AAEOY Most Promising Engineer of the Year

Mr. Wei (Max) Fang grew up in Beijing, where an early interest in photography and videography sparked his curiosity about the technology behind digital imaging. This passion led him to study electronic engineering in Hong Kong before moving to the United States, where he began his career developing system-on-chip technologies powering imaging devices. Over time, Max became an engineer specializing in high-performance computing systems for artificial intelligence and imaging workloads. His work spans hardware architecture, compiler systems, and large-scale machine learning inference, focusing on bridging advanced algorithms with efficient custom hardware execution.

Max has contributed to the architecture and modeling of multiple compute subsystems, including image and video processing pipelines, AI memory fabrics, DMA engines and high-performance data paths. He also developed microcode for image processing components such as temporal noise reduction. Building on this foundation, Max has defined architectural enhancements for essential blocks like memory fabrics and matrix computation cores.

Recognizing that hardware performance depends heavily on software tooling, Max led the development of a compiler framework that translates high-level machine learning workloads, such as ONNX models, into an optimized intermediate representation for custom dataflow hardware. He initially defined the compiler's intermediate representation and progressive lowering strategy, establishing a scalable cross-team interface between model frameworks, optimizers and low-level compiler layers. Within this framework, he also designed and led the development of an automatic calibration-based post-training quantization system, enabling machine learning models to be efficiently deployed on specialized hardware within minimal manual tuning. This capability significantly simplifies model portability for customers and differentiates the platform from many emerging AI accelerator solutions.

Max has also led several engineering teams delivering production systems. His team is responsible for the compiler optimizer framework. At the early stage of the recent wave of large language model adoption, when end-to-end inference pipelines were still rapidly evolving, he led a small team to implement LLM inference on the existing hardware with a whole new programming paradigm in less than half a year. Max continues his leadership in the development of a novel dynamic graph compiler targeting next generation AI SoCs.

As an Asian American engineer working across cultures and disciplines, Max values building technologies that make advanced computing more accessible and impactful.