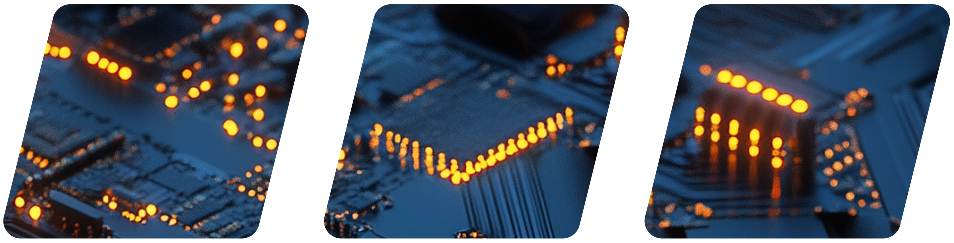
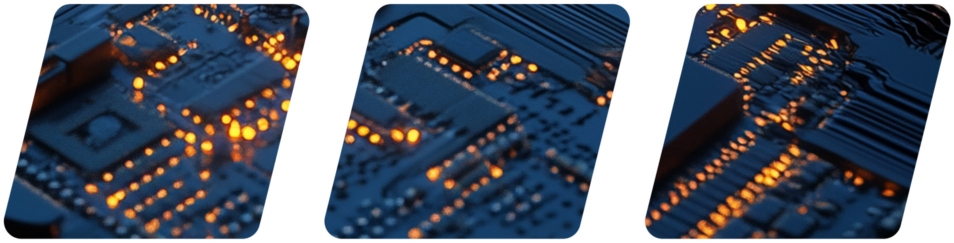


Lattice sensAI™ 8.0 Low Power, Real-time AI for the Far Edge



White Paper

Author:
Deepak Boppana, Senior Director, Product Marketing, Lattice Semiconductor
Hussein Osman, Director, Strategic Business Development, Lattice Semiconductor
Nicolas Widynsky, AI Fellow, Lattice Semiconductor

DISCLAIMERS

Lattice makes no warranty, representation, or guarantee regarding the accuracy of information contained in this document or the suitability of its products for any particular purpose. All information herein is provided AS IS, with all faults, and all associated risk is the responsibility entirely of the Buyer. The information provided herein is for informational purposes only and may contain technical inaccuracies or omissions, and may be otherwise rendered inaccurate for many reasons, and Lattice assumes no obligation to update or otherwise correct or revise this information. Products sold by Lattice have been subject to limited testing and it is the Buyer's responsibility to independently determine the suitability of any products and to test and verify the same. Lattice products and services are not designed, manufactured, or tested for use in life or safety critical systems, hazardous environments, or any other environments requiring fail-safe performance, including any application in which the failure of the product or service could lead to death, personal injury, severe property damage or environmental harm (collectively, "high-risk uses"). Further, buyer must take prudent steps to protect against product and service failures, including providing appropriate redundancies, fail-safe features, and/or shut-down mechanisms. Lattice expressly disclaims any express or implied warranty of fitness of the products or services for high-risk uses. The information provided in this document is proprietary to Lattice Semiconductor, and Lattice reserves the right to make any changes to the information in this document or to any products at any time without notice.

INCLUSIVE LANGUAGE

This document was created consistent with Lattice Semiconductor's inclusive language policy. In some cases, the language in underlying tools and other items may not yet have been updated. Please refer to Lattice's inclusive language FAQ 6878 for a cross reference of terms. Note in some cases such as register names and state names it has been necessary to continue to utilize older terminology for compatibility.

ABSTRACT

As AI workloads shift from the cloud to the far edge, embedded systems must deliver real-time perception within strict power, size, and lifecycle constraints that conventional GPU- and SoC-based architectures struggle to meet. This white paper introduces Lattice sensAI 8.0, a comprehensive far edge AI solution stack that enables advanced computer vision on low power FPGA platforms with deterministic latency and sub-watt power consumption. Lattice sensAI 8.0 combines a redesigned machine learning accelerator, expanded operator and topology support for modern detection networks, a unified production-ready Model Zoo, and deployment-accurate quantization, simulation, and compilation workflows. The platform is complemented by the Golden AI Reference Design (GARD), which accelerates development and reduces deployment risk across automotive, industrial, robotics, smart city, and human–machine interface applications. By aligning modern perception models with the fundamental constraints of embedded systems, Lattice sensAI 8.0 provides a scalable, secure, and lifecycle-aware foundation for always-on, real-time AI at the far edge.

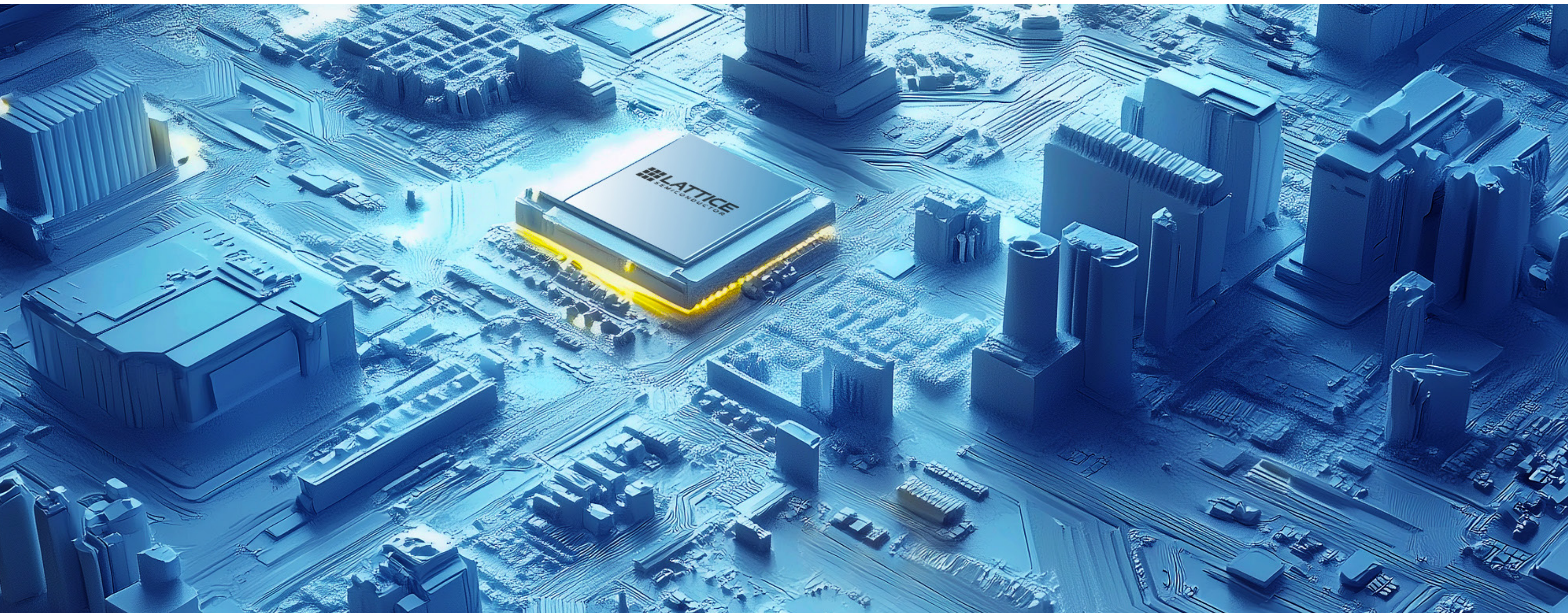


TABLE OF CONTENTS

Disclaimers 2
Inclusive Language 2
Abstract 2
Introduction 4
Why Lattice sensAI 8.0 Matters Now 4
The Limits of Conventional AI Architectures at the Far Edge 4
Market and Technology Trends 4
The Lattice sensAI 8.0 Solution Stack 5
What's New in Lattice sensAI 8.0 6
GARD: A Platform-Level Reference Architecture 8
Benefits and Differentiators 8
Conclusion 10
Appendix 10

Introduction

The Lattice sensAI version 8.0 solution stack makes practical AI inference at the far edge by enabling real-time computer vision within the strict power, size, and reliability constraints of embedded systems. Operating at sub-watt power levels, Lattice sensAI 8.0 supports always-on applications across automotive sensing, industrial inspection, robotics, and human-machine interfaces, use cases that are often impractical with GPU- or SoC-based approaches.

Compared to Lattice sensAI 7.0, the 8.0 release expands support for modern object-detection architectures, introduces a unified, production-ready Model Zoo, and incorporates a redesigned ML accelerator core that delivers significant reductions in logic area on Lattice Nexus™-based devices. These improvements lower system cost while improving scalability and design reuse.

Existing customers benefit from backward compatibility and a streamlined migration path, while new users can accelerate development using GARD reference platforms and pre-validated models.

Strategically, Lattice sensAI 8.0 positions Lattice as a flexible far-edge AI platform, reducing customer time-to-market while creating a durable foundation for long-lifecycle, power-constrained edge systems.

Why Lattice sensAI 8.0 Matters Now

AI workloads are increasingly moving from centralized cloud infrastructure to the edge, driven by the need for lower latency, deterministic behavior, reduced power consumption, and data privacy. Vision-based systems in industrial, automotive, and human–machine interface applications must now operate continuously within sub-watt power envelopes while delivering real-time perception in thermally and mechanically constrained environments. These requirements expose fundamental limitations in traditional GPU- and SoC-centric AI approaches.

Lattice sensAI 8.0 addresses this inflection point by enabling advanced computer vision networks, including modern object detection architectures such as YOLOv8 and YOLOv11, on compact, low power FPGA platforms. Through a redesigned machine-learning accelerator, a unified, production-ready Model Zoo, and an integrated development toolchain, sensAI 8.0 makes sophisticated perception feasible at the far edge.

The Limits of Conventional AI Architectures at the Far Edge

Most computer vision AI frameworks are optimized for datacenters or high-end SoCs, rather than the resource-limited environments that define embedded and industrial systems. These constraints include:

- **Strict power and thermal limits** – Always-on vision workloads must operate at only a few hundred milliwatts.
- **Latency sensitivity** – Safety-critical systems require consistent real-time inference.
- **Extended lifespan and flexibility** – Products may need upgrades or new features deployed in the field over many years.
- **Computational restrictions** – Edge devices lack the compute and memory usually required by modern CNNs and transformers.

Existing solutions fall short. NPUs deliver performance but offer limited configurability and deterministic control. Fixed-function ASICs cannot evolve as algorithms change. Embedded CPUs struggle to meet real-time requirements without exceeding power budgets.

While FPGAs address many of these architectural challenges, they have historically lacked accessible tools for efficiently deploying modern neural networks. This gap between far-edge system constraints and practical AI deployment motivated the design of Lattice sensAI 8.0.

Market and Technology Trends

Edge AI adoption is rapidly accelerating, with the global embedded and edge AI hardware market projected to grow from approximately $17.8 billion in 2024 to over $46.4 billion by 2029 (VDC Research, “Embedded & Edge AI Hardware,” September 2025). This growth is driven by applications that demand real-time perception under strict power, cost, and deployment constraints. See Table 1.

Table 1: Key Sectors Driving Market Growth

Key Sectors Driving This Growth	
Automotive	Driver monitoring systems (DMS), occupant monitoring systems (OMS), external monitoring systems (EMS), camera management systems (CMS), and Advanced Driver Assistance Systems (ADAS) all demand real-time perception within strict power and cost envelopes.
Industrial Automation	Defect detection, machine vision, and predictive maintenance depend on continuous inspection and anomaly detection in harsh environments.
Smart Cities	Traffic flow analysis, vehicle and pedestrian classification, and infrastructure monitoring demand distributed intelligence without steady cloud dependency.
Robotics	Obstacle avoidance, multi-object perception, mapping, and navigation require low-latency, power-efficient vision.
Human–machine Interfaces (HMI)	Context-aware interfaces, presence detection, and gesture tracking depend on low power, always-on sensing.

Across these markets, the common requirement is deterministic, low power vision intelligence deployed close to the sensor, a combination that exposes the limitations of conventional AI accelerators. These trends align directly with Lattice's low power FPGA architecture, positioning Lattice sensAI 8.0 to address a growing class of edge AI applications that demand both flexibility and efficiency.

The Lattice sensAI 8.0 Solution Stack

The Lattice sensAI 8.0 solution stack delivers an end-to-end platform for developing, deploying, and maintaining edge AI models on Lattice FPGAs. Designed for the far edge, it addresses real-time, low power AI requirements across automotive, industrial, and smart city applications.

Key Components

Lattice sensAI 8.0 natively integrates hardware and software to meet the real-time, low power, and long-lifecycle demands of far-edge systems. By unifying perception models, development tools, and hardware acceleration into a single platform, it is purpose-built for embedded and industrial deployments on Lattice Semiconductor FPGAs.

At its core, Lattice sensAI 8.0 combines perception models, deployment-ready tools, and a resource-efficient machine-learning accelerator. This co-designed approach enables deterministic performance, rapid iteration, and low power consumption, reducing both time-to-market and total cost of ownership. Together, these elements form a unified far-edge AI platform that enables sophisticated perception in power- and resource-constrained environments where conventional AI approaches fall short. The key components are:

- **Purpose-Built Models** – A unified repository of models for HMI, automotive, industrial, and robotics workloads. Each model includes a model card, training scripts, and configuration files.
- **Deployment-Accurate Development Flow** – Integrated training and quantization tools support Fixed-Point Quantization (FPQ), Learned Step-Size Quantization (LSQ), and Post-Training Quantization (PTQ), enabling developers to validate accuracy and operation under hardware-realistic conditions.
- **Pre-Deployment Verification and Optimization** – A bit-accurate ML Engine Simulator mirrors on-device behavior, allowing developers to debug models, assess latency and throughput, and maximize resource usage before committing hardware.
- **Hardware-Aware Compilation and Acceleration** – The neural network compiler maps high-level model descriptions into optimized configurations for Lattice ML IP, managing scheduling, tiling, and memory allocation. A redesigned CNN accelerator core delivers significant area reductions while retaining performance and accuracy.
- **Validated Reference Platforms** – GARD, Lattice CertusPro^{TM}-NX System-on-Module (SoM), and Lattice CrossLink^{TM}-NX-based camera boards deliver validated hardware and software solutions for rapid prototyping and deployment across automotive and industrial domains.

This release delivers meaningful advances, including substantial reductions in LUT and register area on Lattice Nexus-based hardware, while maintaining performance. See Figure 1 for the supported Lattice FPGA product families.

Figure 1: Supported Lattice FPGA Product Families

What's New in Lattice sensAI 8.0

Lattice sensAI 8.0 represents a substantive evolution of the edge AI stack, extending model compatibility, improving deployment efficiency, and substantially lowering FPGA resource utilization. Building on the operator coverage and workflows introduced in sensAI 7.0, it expands support for modern, anchor-free detection architectures and production-scale deployments.

Where Lattice sensAI 7.0 focused on forming a resilient far-edge AI foundation, Lattice sensAI 8.0 shifts the platform toward scalability, architectural headroom, and long-term maintainability.

Key advances include:

- Broader operator and topology support for state-of-the-art vision models
- A unified, production-ready Model Zoo
- A redesigned ML accelerator with substantial area savings
- More explicit, deployment-accurate quantization workflows
- YAML-based topology descriptions to support automation and governance

Expanded Topology and Operator Support

Lattice sensAI 8.0 expands support for modern neural network operations required by current and emerging computer vision architectures. New operators such as SiLU, UpSampling2D, Split, K-MaxPooling, MatMul, Softmax, and GlobalAveragePooling2D (GAP2D) enable more expressive network designs and multi-scale detection pipelines.

These additions allow sensAI 8.0 to efficiently deploy anchor-free detectors and deeper perception networks while maintaining deterministic execution and hardware efficiency. Enhancements to the address generator and ML IP scheduling improve support for complex detection heads and modern activation functions.

Unified and Production-Ready Model Zoo

Lattice sensAI 8.0 unifies the Model Zoo with purpose-built models optimized for deployment rather than experimentation. Models are delivered as standardized, version-controlled artifacts with consistent documentation, configuration, and training assets.

Each Model Zoo entry includes:

- A structured model card
- Training and fine-tuning scripts
- Dataset and preprocessing references
- Deployment configuration files
- Integration with GARD reference platforms and the sensAI toolchain

This method decreases integration risk, simplifies validation, and enables consistent reuse across projects and product generations.

HMI and Contextual Vision Models

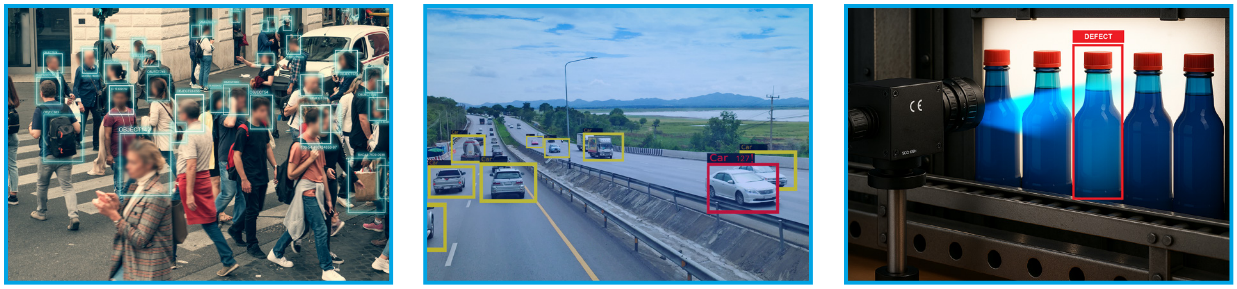
The Model Zoo includes a set of perception models optimized for always-on, low power contextual awareness, including person detection, face detection and identification, facial landmarks, and hand detection and tracking.

These models are designed for applications where deterministic latency, privacy, and power efficiency are critical, such as presence detection, gesture-based interaction, driver and occupant monitoring, and safety-related perception. See Figure 2.

For example, these validated models enable always-on presence detection at approximately 15 mW in ultra-low power standby modes. Furthermore, they deliver the deterministic responsiveness required for interactive systems, with inference latencies as low as 25 ms for face detection and 28 ms for facial landmark validation when deployed on Lattice Nexus-based FPGAs.

By delivering these capabilities as validated, deployable models, Lattice sensAI 8.0 enables contextual vision to function as a standard system capability rather than a custom implementation. As a result, sensAI 8.0 extends beyond generic vision tasks to support application-ready perception systems.

Figure 2: High-growth Edge AI Workloads



Generic Multi-Object Detection

COCO-based multi-class detector suitable for robotics, smart-city analytics, and domain-specific fine-tuning.

Automotive Multi-Object Detection

An anchor-free detector optimized for automotive-relevant object classes and real-time perception under strict power constraints.

Few-Shot Defect Detection

Siamese-based architecture for industrial inspection, supporting fast modification for new defect types using restricted tagged datasets.

ML Accelerator Improvements: Area Efficiency and Architectural Headroom

Lattice sensAI 8.0 introduces a redesigned convolution engine in the ML accelerator that substantially reduces FPGA logic utilization while retaining latency, numerical behavior, and accuracy. Across representative configurations, logic usage is reduced by approximately 30–40 percent, creating additional architectural headroom for deeper networks, multi-scale detection, or added system functionality.

The updated accelerator architecture maintains consistent DSP and memory utilization, supports matrix-multiplication-centric operations, and improves streaming efficiency through a new address generator. These enhancements allow Lattice sensAI 8.0 to scale with increasing model complexity without increasing power consumption or compromising determinism.

Quantization in Lattice sensAI 8.0

Quantization is an essential part of Lattice sensAI 8.0, enabling advanced neural networks to run efficiently on resource-constrained FPGAs. Lattice sensAI 8.0 supports three complementary quantization approaches, each serving a particular role in the development flow.

- **Fixed-Point Quantization (FPQ)** provides deterministic, deployment-accurate inference that corresponds exactly to hardware behavior.
- **Learned Step-Size Quantization (LSQ/QAT)** improves accuracy by jointly training quantization parameters with model weights, particularly for deeper networks and non-linear activations.
- **Post-Training Quantization (PTQ)** supports rapid feasibility analysis and legacy adaptation where retraining is impractical.

By explicitly defining the role of each method, Lattice sensAI 8.0 improves alignment between model training, simulation, and deployment.

YAML-Based Topology Definitions

Lattice sensAI 8.0 introduces YAML-based topology descriptions as a first-class artifact in the development workflow. YAML files capture network structure, quantization parameters, preprocessing stages, and target hardware configuration in a compact, human-readable format.

This solution supports:

- A single source of truth across ML, FPGA, and firmware teams
- Integration with CI/CD pipelines for retraining, compilation, and regression testing
- Improved auditability through versioned, diffable specifications
- Faster debugging and support through shareable configuration artifacts

YAML-based descriptions establish a scalable foundation for repeatable, governable, and automated edge AI deployment, aligning sensAI 8.0 with modern software-hardware co-development practices.

GARD: A Platform-level Reference Architecture

The GARD is a platform-level implementation that integrates the Lattice sensAI solution stack into a hardware-and-firmware architecture. It provides:

- **Camera and ISP Pipeline** – Handling sensor input, demosaicing, basic image signal processing, and scaling.
- **Memory Subsystem** – Including TCM, HyperRAM, and execute-in-place (XIP) Flash options.
- **ML Subsystem** – Hosting the ML IP core, associated buffers, and DMA engines.
- **RISC-V Subsystem** – Executing pre-processing, post-processing, control, and application logic.
- **Host Integration (HUB)** – Providing streaming, visualization, telemetry, and configuration interfaces for development and evaluation.

The GARD accelerates time-to-value by providing validated reference firmware, application examples, and host tools that demonstrate sensAI 8.0 models in real-world applications, such as multi-object detection and defect detection.

Benefits and Differentiators

Lattice sensAI 8.0 delivers differentiated value across technical performance, business outcomes, security, and developer productivity, enabling scalable, sustainable edge AI deployments. Key benefits and differentitators are summarized in Table 2. These features streamline integration with CI/CD pipelines and help maintain repeatable, scalable edge AI workflows.

Key Benefits:

- **Technical** – Deterministic, ultra-low power AI for advanced vision models
- **Business** – Lower system cost, scalable platforms, and long product lifecycles
- **Security** – Secure boot, encrypted models, and trusted deployment
- **Developer** – Automated, reproducible workflows with hardware-accurate validation

Table 2: Lattice sensAI 8.0 Benefits and Differentiators

Technical Benefits	
Ultra-low power with deterministic latency	Supports always-on, real-time inference for safety- and interaction-critical embedded systems
Advanced model support at the far edge	Enables deployment of YOLOv8- and YOLOv11-class workloads inside compact FPGA footprints
Unified ML IP architecture	Consistent behavior and performance across Lattice CertusPro-NX and Lattice Avant device families
Purpose-Built Model Zoo	Validated models spanning HMI, automotive, industrial, and robotics use cases

Business and Product Benefits	
Lower system cost and BOM efficiency	Low power FPGA-based designs reduce component count and thermal complexity, enabling compact products with predictable lifecycle costs
Platform reuse and SKU scalability	A single hardware and software foundation can support multiple product variants and feature sets, improve reuse, and accelerate time-to-market
Field upgradability over long lifecycles	Secure updates to models and firmware enable continuous improvement, regulatory adaptation, and extended product relevance
Enterprise-grade security	Secure boot, encrypted model delivery, and remote attestation protect IP, data, and system integrity across deployment environments

Developer Experience	
End-to-end AI workflow	From training through quantization, simulation, compilation, and deployment
Automation-friendly APIs	YAML and Python-based interfaces enable scripting, reproducibility, and workflow integration
Robust simulation and debugging	Bit-accurate simulation and profiling reduce iteration time and deployment risk

CI/CD and MLOps Integration Lattice sensAI 8.0 aligns with modern automation and MLOps practices, supporting:
Scriptable CLI and Python APIs for automated model training, evaluation, and deployment
Bit-accurate simulation for validation and regression testing
Environment management and packaging with Conda and standard distribution formats
Automated benchmarking to monitor system performance
Secure lifecycle management, including versioned updates and remote deployment

Conclusion

The Lattice sensAI 8.0 solution stack makes advanced AI inference practical at the far edge by aligning modern perception models with the core limitations of embedded systems. Through a reduced-footprint ML accelerator, expanded operator and model support, a unified Model Zoo, and an integrated, configurable toolchain, Lattice sensAI 8.0 enables intelligent, always-on systems that operate within strict power, cost, and size limits.

Lattice sensAI 8.0 represents a shift from point solutions toward a scalable edge AI platform. By combining deterministic hardware acceleration with deployment-ready software workflows, Lattice enables customers to design systems that can evolve, supporting new models, features, and regulatory requirements without architectural redesign.

As AI moves closer to the sensor, success at the far edge will depend less on raw compute and more on efficiency, reliability, and lifecycle adaptability. Lattice sensAI 8.0 positions Lattice Semiconductor as a key enabler of this transition, providing a foundation for the next generation of low power, real-time, and context-aware edge intelligence.

Appendix

Supported Operating Systems and Toolchains

Lattice sensAI 8.0 is designed to integrate into a wide range of development and deployment environments commonly used for embedded and edge AI systems. The platform supports both Windows- and Linux-based workflows and aligns with standard Python-based machine learning ecosystems.

Supported Operating Systems

Lattice sensAI 8.0 development tools and workflows are supported on the following host operating systems:

- **Linux (x86_64)** – Primary development environment for training, quantization, simulation, and compilation workflows.
- **Microsoft Windows (x86_64)** – Supported for development workflows, including tool execution and model preparation. Linux-based components may be accessed via compatible subsystems or virtualized environments where applicable.

Target deployment environments typically include embedded Linux systems running on host processors paired with Lattice FPGAs.

Toolchain and Environment Support

Lattice sensAI 8.0 integrates with standard machine learning and software development tools to support automated and reproducible workflows:

- **Python-based workflows** – Used for model training, quantization, evaluation, and automation.
- **Conda environments** – Recommended for dependency management and environment consistency across development, testing, and production systems.
- **Command-line interfaces (CLI)** – Enable scripted execution of quantization, simulation, compilation, and deployment steps.
- **Package-based deployment artifacts** – Support for .whl and .deb formats enables consistent distribution and installation across systems.

Integration with Development Pipelines

The sensAI 8.0 toolchain is designed to support integration with CI/CD and automated development pipelines through:

- Scriptable tool execution
- Version-controlled configuration artifacts (e.g., YAML-based topology descriptions)
- Hardware-accurate simulation for pre-deployment validation
- Structured packaging and artifact management

These capabilities enable teams to adopt sensAI 8.0 within existing software and hardware development practices without requiring custom infrastructure.

READY TO LEARN MORE?

To learn more about Lattice low power FPGA-based solutions for industrial, automotive, communications, computing, and consumer applications, visit **www.latticesemi.com** or contact us at **www.latticesemi.com/contact** or **www.latticesemi.com/buy**.

TECHNICAL SUPPORT ASSISTANCE

Submit a technical support case through **www.latticesemi.com/techsupport**.
For frequently asked questions, please refer to the Lattice Answer Database at
www.latticesemi.com/Support/AnswerDatabase.