QADS AI EDGE SoC PLATFORM

ENGINEERING REFERENCE

Modular Scalable ASIC Family Architecture — Nano / Lite / Pro / Max

qads.ledcircle.com
zeno@ledcircle.com

Prepared for prospective design partners and customer engineering teams.

Contents

1. Platform Overview

QADS AI Edge SoC Platform is a scalable, modular silicon architecture for battery-powered Edge AI applications, spanning ultra-low-power IoT sensing through high-performance edge AI compute on mature-to-advanced process nodes.

This reference document is intended for customer engineering teams evaluating QADS as a derivative-design platform. It covers the architecture, tier specifications, IP content, software stack, and engagement model relevant to a technical evaluation.

1.1 Platform Goals

- Enable rapid SoC derivative development from a shared architecture
- Reduce ASIC re-design cost across a product line
- Support AI-first edge computing workloads
- Provide a unified software and hardware ecosystem
- Scale from ultra-low-power IoT to high-performance edge AI within a coherent platform

2. Architecture Concept

2.1 Modular Silicon Block Architecture

The platform is organized around a central platform core and six modular tiles, connected through a shared NoC/AXI fabric.

CPU Tile	AI Tile	Video Tile
Cortex-M / RISC-V	NPU (0.5–16 TOPS)	ISP + CODEC

Memory Tile	Connectivity	Security Tile
DDR / SRAM	Wi-Fi / BLE	Root of Trust

All tiles interconnect through the QADS NoC / AXI Fabric (Section 5).

3. Product Family & Tier Specifications

Four tiers span the platform's target market, sharing a common architecture and scaling CPU, AI compute, memory, and power envelope per tier.

Tier	Target Use Case	CPU	AI NPU	Memory	Power	Indicative Node
QADS Nano	Sensors / IoT	M0+	0.2–0.5 TOPS	32–64MB	<100mW	22/28nm
QADS Lite	Smart Camera	M33	0.5–1 TOPS	128MB	<500mW	22/16nm
QADS Pro	IPC / Vision AI	M33/M55	2–4 TOPS	256MB	<1.5W	12/16nm
QADS Max	Edge AI Box	RISC-V / M7	8–16 TOPS	512MB+	<5W	7/12nm

Process node figures are indicative targets for engineering planning purposes and should be confirmed against final foundry agreements during detailed design.

4. Core Platform IP Blocks

4.1 Compute Cluster

- ARM Cortex-M33 / RISC-V RV32/64
- Optional DSP extension
- TrustZone support
- DVFS engine
- Interrupt controller (GIC-lite)

4.2 AI Engine (QADS-NPU)

- Scalable 0.2 → 16 TOPS architecture
- INT8 / INT16 / FP16 support
- Layer fusion engine
- Sparse computation support
- On-chip tensor scheduler
- Hardware model decomposer

4.3 Video Processing Unit

- MIPI CSI-2 RX
- ISP pipeline (HDR / NR / AWB / AE)
- JPEG / H.264 / H.265 encoder
- Frame buffer compression
- Low-latency streaming mode

4.4 Memory System

- LPDDR / DDR SiP architecture
- SRAM scratchpad

- Cache hierarchy (L1/L2 optional)
- DMA controller
- AXI / NoC interconnect

4.5 Connectivity Subsystem

- Wi-Fi 4/5/6 scalable MAC
- BLE 5.x
- Optional Ethernet MAC
- USB 2.0 (optional tier)

4.6 Security Subsystem (Root of Trust)

- Secure Boot ROM
- Hardware Root Key
- AES-256 engine
- SHA-2 engine
- ECC / RSA acceleration
- TRNG
- eFuse / OTP
- Secure OTA rollback protection

5. Platform Interconnect (Q-NoC)

5.1 Architecture

- AXI4-based scalable NoC
- QoS arbitration engine
- Priority-based bandwidth allocation

5.2 Traffic Classes

- Real-time video stream
- AI inference data
- CPU control traffic
- Network streaming traffic

6. Power Architecture

6.1 Power Domains

- Always-On Domain (AON)
- Compute Domain
- AI Domain
- Multimedia Domain
- RF Domain

6.2 Power States

State	Description	Power
Deep Sleep	AON only	<150 µA
Wake	Sensor trigger	<10 mA
AI Active	NPU running	50–300 mA
Full System	AI + Wi-Fi TX Peak	1–1.5W

7. Software Platform

- Application Layer
- AI SDK (QADS AI Runtime)
- TensorFlow Lite / ONNX Runtime
- Media Framework
- FreeRTOS / RTOS Kernel
- HAL / Driver Layer
- QADS Silicon Platform

7.1 AI Runtime

- Model compiler
- Quantization toolchain
- Operator fusion
- Memory optimizer
- NPU scheduler

7.2 Cloud Connectivity

- MQTT
- HTTPS
- OTA engine
- Device management
- Fleet management API

8. Cross-Tier Software Compatibility

The NPU is defined as a single ISA family with a common operator set; higher tiers are supersets of lower tiers.

Tier	Operator Set	Precision	Compiler Target
Nano	Base set (conv, pooling, activation, FC)	INT8	Common IR — Base
Lite	Base + depthwise / grouped conv	INT8 / INT16	Common IR — Base+
Pro	Base+ + attention / transformer ops	INT8 / INT16 / FP16	Common IR — Full
Max	Full set + sparse / dynamic shape	INT8 / INT16 / FP16	Common IR — Full

8.1 Toolchain Notes

- Single model compiler front-end across all four tiers; tier selection is a build target flag

- Quantization toolchain shares a calibration pipeline across tiers; only output precision differs
- ONNX / TFLite import path is identical across tiers; divergence appears only at operator-lowering
- A model built against the Pro/Max operator set is not guaranteed to run unmodified on Nano/Lite; the compiler flags unsupported operators for the target tier at build time

9. Security & Certification Roadmap

Certification	Relevance	Target Tier(s)	Status
PSA Certified (Level 1–2)	Standard expectation for IoT/edge silicon buyers	Nano, Lite, Pro	Target
FIPS 140-3 (crypto module)	Industrial / government-adjacent customers	Pro, Max	Target
Common Criteria (EAL2+)	Security-sensitive verticals	Max	Future consideration

10. Engagement & Delivery Models

QADS supports multiple engagement structures depending on customer design capability and program scope.

Model	How It Works	Best Fit
IP Licensing	QADS licenses platform IP (tiles, NoC, AI runtime); the customer's own design team takes the derivative to tape-out	Customers with an internal design team who want speed, not outsourced design
Hybrid (IP + Design Service)	QADS licenses platform IP and also provides paid derivative design services to customize and tape out the derivative on the customer's behalf	Customers who want a derivative chip but lack in-house RTL/physical design capacity
Alliance-Built Integration	QADS integrates partner or licensed IP blocks to accelerate a specific derivative where deeper customization is required	Programs needing IP outside the standard platform tile set

- GDS-ready IP blocks
- Standard AXI interface
- Pre-verified subsystem blocks

11. Platform Scalability

11.1 Silicon Reuse Model

- High IP reuse expected across tiers sharing the same process node
- Modular tile-based architecture
- Drop-in NPU scaling across the product family

11.2 Design Philosophy

- AI-first silicon architecture
- Power-first design constraint
- Modular scalability over monolithic SoC
- Edge autonomy (no-cloud dependency)
- Security-by-design

12. Contact

For technical evaluation, derivative scoping, or engagement discussions, please reach out directly.

Web qads.ledcircle.com
Email zeno@ledcircle.com

QADS — QUANTUM AI DESIGN SERVICES