QADS

EDGE AI PLATFORM

LEADING-EDGE SILICON FOR HIGH-PERFORMANCE EDGE AI

2nm – 5nm Platform Family · v1.0

Vision AI · Advanced Robotics · On-Device Generative AI · Autonomous Systems

qads.ledcircle.com

zeno@ledcircle.com

Platform Vision

The QADS Edge AI Platform is a leading-edge, high-performance silicon architecture built for battery- and power-constrained Edge AI systems that demand data-center-class inference at the edge — vision AI boxes, advanced robotics, on-device generative AI, and autonomous systems.

The platform delivers:

- Data-center-class AI inference performance within edge power and thermal envelopes
- A focused 2nm–5nm leading-edge node range, purpose-built for the highest-performance tier of edge AI
- Rapid derivative development across a tightly integrated tier family
- A unified software and hardware ecosystem built for transformer-class and generative edge AI workloads
- Design and delivery through Qoresic's AI-native IC design service

Why Leading-Edge Nodes (2nm–5nm)

Edge deployment of transformer-based and generative models requires substantially more compute than earlier-generation CNN-era edge AI workloads. Sub-7nm nodes are required to deliver the necessary TOPS within real-world edge power envelopes.

QADS concentrates exclusively on this high-performance band, giving customers a platform purpose-built for leading-edge performance — rather than a general-purpose architecture stretched across many power tiers.

Platform Architecture

Modular Silicon Block Architecture

The platform is built on a modular tile architecture, re-specified for leading-edge performance targets:

Tile	Capability
CPU Tile	High-performance Arm / RISC-V cluster
AI Tile	NPU, 16–128 TOPS
Video / Sensor Tile	Multi-sensor ISP + CODEC + fusion
Memory Tile	LPDDR5X / HBM-lite + SRAM
Connectivity Tile	Wi-Fi 6E/7, 5G option, high-speed SerDes

Tile	Capability
Security Tile	Root of Trust + secure enclave

All tiles interconnect through the QADS NoC / die-to-die fabric.

Scalable Product Family

Three tiers span the leading-edge band, each tuned for a distinct class of edge AI deployment:

Tier	Target Use Case	AI NPU	Memory	Power	Node
QADS Edge-5	Smart camera / robotics perception	16–32 TOPS	8–16GB LPDDR5X	3–6W	5nm
QADS Edge-3	On-device gen-AI, advanced ADAS	32–64 TOPS	16–32GB LPDDR5X / HBM-lite	6–12W	3nm
QADS Edge-2	Autonomous systems, edge AI server-in-a-box	64–128 TOPS	32GB+ HBM-lite	12–25W	2nm

Core Platform IP Blocks

Compute Cluster

- High-performance Arm Cortex-A / Neoverse-class or RISC-V RVA cluster
- Multi-core coherent cache hierarchy
- TrustZone / secure world support
- Advanced DVFS with per-core power islands

AI Engine (QADS-NPU, Leading-Edge Variant)

- Scalable 16 → 128 TOPS architecture
- INT4 / INT8 / FP8 / FP16 support for transformer and generative workloads
- Native attention / transformer operator acceleration
- Sparse and structured-sparse computation support
- On-chip tensor scheduler with dynamic shape support
- Multi-die NPU scaling for the Edge-2 tier

Video / Sensor Fusion Unit

- Multi-camera MIPI CSI-2 RX (up to 8 streams)
- ISP pipeline (HDR / NR / AWB / AE)
- AV1 / H.265 encode-decode

- Sensor fusion engine (camera + LiDAR/radar input paths)
- Low-latency streaming mode

Memory System

- LPDDR5X primary memory; HBM-lite option for Edge-2
- Large SRAM scratchpad for on-chip model weight caching
- Multi-level cache hierarchy
- High-bandwidth DMA + NoC interconnect

Connectivity Subsystem

- Wi-Fi 6E/7 MAC
- 5G modem option (Edge-2 tier)
- High-speed SerDes for sensor/accelerator expansion
- Automotive-grade Ethernet option

Security Subsystem (Root of Trust)

- Secure Boot ROM
- Hardware Root Key + secure enclave
- AES-256 / SHA-3 engines
- Post-quantum-ready crypto accelerator slot
- TRNG, eFuse/OTP
- Secure OTA with rollback protection

Platform Interconnect (Q-NoC)

Architecture

- AXI5/CHI-based coherent NoC
- Die-to-die interconnect for Edge-2 multi-die configurations
- QoS arbitration with AI-workload priority class

Traffic Classes

- AI inference data (highest priority)
- Multi-stream video / sensor fusion data
- CPU control traffic
- Network / connectivity traffic

Power Architecture

Power Domains

- Always-On Domain (AON)
- Compute Domain
- AI Domain (independently scalable)
- Multimedia / Sensor Domain
- Connectivity Domain

Power States

State	Description	Power
Idle	AON + display only	<300 mW
Standby	Sensor monitoring active	<1 W
AI Active	NPU running inference	2–15 W (tier-dependent)
Full System	AI + sensor fusion + connectivity peak	6–25 W (tier-dependent)

Software Platform (QADS OS Stack)

- Application Layer
- AI SDK — QADS AI Runtime, optimized for transformer and generative models
- PyTorch / ONNX Runtime / GGUF-class quantized model support
- Media + Sensor Fusion Framework

- Linux / RTOS hybrid kernel option
- HAL / Driver Layer
- QADS Silicon Platform

AI Runtime

- Model compiler with transformer/attention operator support
- INT4/INT8/FP8 quantization toolchain
- KV-cache optimization for on-device generative inference
- Multi-die NPU scheduler (Edge-2 tier)

Cloud Connectivity

- MQTT / HTTPS
- OTA engine
- Fleet management API
- Optional hybrid edge-cloud inference offload

Security & Certification Roadmap

Certification	Relevance	Target Tier(s)
PSA Certified (Level 2–3)	Baseline for edge AI silicon at this performance tier	Edge-5, Edge-3, Edge-2
FIPS 140-3	Industrial / government-adjacent and automotive-adjacent deployments	Edge-3, Edge-2
ISO 26262 (ASIL-B/D readiness)	Prerequisite for ADAS / autonomous-systems customers	Edge-2 (automotive variant)
Common Criteria (EAL4+)	Differentiator for security-sensitive robotics / defense-adjacent deployments	Edge-2

Designed by Qoresic, Owned by QADS

QADS is the platform brand and product owner. QADS defines requirements, owns the platform roadmap and tier strategy, and is the commercial face of the Edge AI Platform to customers.

Qoresic is QADS's AI-native IC design service provider, executing the full design flow — specification through tape-out — on QADS's behalf, using its AI Supervisor Brain and agent teams spanning architecture, RTL, verification, DFT, physical design, and sign-off.

Platform Scalability

A narrow, single-band node range (2–5nm) drives high IP reuse across the Edge-5, Edge-3, and Edge-2 tiers. The modular tile-based architecture scales from single-die configurations up to multi-die NPU scaling reserved for the Edge-2 tier.

The Edge AI Platform is a higher-performance sibling to the broader QADS silicon family. Customers needing ultra-low-power, sub-1W operation are served by other QADS platform tiers; customers needing leading-edge generative and transformer-class edge inference are served by the Edge AI Platform.

Design Philosophy

- Performance-first within edge power constraints
- AI-native silicon architecture, optimized for transformer/generative workloads
- Modular scalability within a single leading-edge node band
- Edge autonomy with optional hybrid cloud offload
- Security-by-design

Why QADS

- A platform purpose-built for the highest-performance tier of edge AI — not a scaled-down data-center chip
- A unified hardware and software stack tuned for transformer and generative workloads from day one
- An AI-native design flow that enables fast, customer-specific derivatives
- A security-by-design architecture with a clear path to industrial, automotive, and defense-adjacent certification

QADS Edge AI Platform is not a smaller version of a data-center chip.

It is a leading-edge silicon platform purpose-built for high-performance edge intelligence.

Get in Touch

To explore how the QADS Edge AI Platform can fit your next product generation, reach out to our team:

Website: qads.ledcircle.com
Email: zeno@ledcircle.com