PPI – Platvoet Performance Intelligence

WHITEPAPER

Thin Edge & AIoT

KI am Edge in der industriellen Praxis

Architektur, Anwendungsfälle und strategische Bewertung

KI am Edge ist kein Proof-of-Concept mehr — sie ist industrielle Realität. Der globale Edge-AI-Markt wächst von 25 Milliarden USD (2025) auf prognostizierte 120 Milliarden USD (2033). Für Industrieunternehmen entscheidet Edge-KI über Reaktionszeiten in Millisekunden, Betriebskosten und Wettbewerbsfähigkeit. Dieses Whitepaper beschreibt die technologische Architektur, das thin-edge.io-Framework als industriellen Enabler, konkrete AIoT-Anwendungsfälle und einen praxisorientierten Bewertungsrahmen für den Einstieg.

Autor	Ralf Platvoet, Diplom-Ökonom
Organisation	PPI – Platvoet Performance Intelligence
Erscheinungsjahr	2026
Umfang	ca. 22 Seiten
Themenbereich	Industrial IoT · AIoT · Edge Computing · Industry 4.0 / 5.0

platvoet.org

Executive Summary

Die Konvergenz von Künstlicher Intelligenz und Industrial Internet of Things — kurz **AIoT** — verändert die Grundlogik industrieller Datenverarbeitung. Während klassische IIoT-Architekturen Rohdaten in die Cloud übertragen und dort verarbeiten, verlagert AIoT die KI-Inferenz direkt an den Ort der Datenentstehung: auf Maschinenebene, in Produktionshallen, auf mobile Geräte.

Der Schlüsselbegriff dabei ist **Thin Edge**: schlanke, ressourcenarme Edge-Agenten, die auf industriellen Geräten und Gateways laufen und dennoch komplexe Aufgaben ausführen können — von der Protokollübersetzung (OPC-UA/Modbus → MQTT) bis zur lokalen KI-Inferenz. Das Open-Source-Framework **thin-edge.io**, ursprünglich von Cumulocity/Software AG entwickelt und heute ein Eclipse-Foundation-Projekt, hat sich als industrieller Standard für genau diese Aufgabe etabliert.

Für Industrieunternehmen stellen sich heute drei strategische Fragen: Welche Verarbeitungsaufgaben gehören an den Edge — und welche in die Cloud? Welche technologischen Bausteine ermöglichen industrietaugliche KI am Edge? Und wie lässt sich der wirtschaftliche Nutzen quantifizieren, bevor ein Pilotprojekt gestartet wird?

Dieses Whitepaper gibt Antworten auf alle drei Fragen — aus technischer und betriebswirtschaftlicher Perspektive.

Kernaussagen

- Der globale Edge-AI-Markt wächst von ~25 Mrd. USD (2025) auf ~120 Mrd. USD (2033) — Industrie ist der größte Wachstumstreiber.
- Thin Edge (thin-edge.io) ist das führende Open-Source-Framework für ressourcenarme industrielle Edge-Agenten.
- AIoT-Latenz am Edge: < 10 ms für lokale Inferenz — vs. 100–500 ms für Cloud-Roundtrip-Verarbeitung.
- Predictive Maintenance, Qualitätssicherung und Anomalieerkennung sind die drei reifsten AIoT-Anwendungsfälle.
- TinyML ermöglicht KI-Inferenz auf Mikrocontrollern mit < 1 MB RAM — neue Hardware-Generation ab 2025.

1. Begriffe und technologischer Kontext

1.1 IIoT, AIoT und Edge Computing: Das Begriffsdreieck

Drei Begriffe, die häufig synonym verwendet werden — aber unterschiedliche technologische Schichten beschreiben:

IIoT	▸ Industrial Internet of Things: Vernetzung physischer Maschinen, Sensoren und Aktoren in industriellen Umgebungen ▸ Fokus: Konnektivität, Datenerfassung, Fernsteuerung, Predictive Maintenance ▸ Protocols: OPC-UA, Modbus, MQTT, LoRaWAN, PROFINET ▸ Herausforderung: Datenmengen, Latenz, OT/IT-Integration
Edge Computing	▸ Verlagerung von Rechenlast und Entscheidungslogik vom Cloud-Rechenzentrum an den Netzwerkrand ▸ Ziel: Latenzreduktion, Bandbreiteneinsparung, Offline-Fähigkeit, Datensouveränität ▸ Thin Edge: Schlanke Agenten auf ressourcenarmen Geräten (< 100 MB RAM, ARM-Architektur) ▸ Thick Edge: Leistungsstarke Edge-Server mit GPU-Beschleunigung (NVIDIA Jetson, Intel OpenVINO)
AIoT	▸ Artificial Intelligence of Things: Kombination aus IIoT-Konnektivität und KI-Inferenz am Edge ▸ Ermöglicht: lokale Entscheidungen in Echtzeit ohne Cloud-Latenz ▸ KI-Methoden: TinyML, ONNX-Runtime, Edge Impulse, federated Learning ▸ Industry 5.0: KI als Kollaborationspartner menschlicher Operatoren — nicht als Ersatz

1.2 Warum Edge statt (nur) Cloud?

Die Frage ist nicht Cloud oder Edge — sie ist: **Welche Verarbeitungsaufgaben gehören wohin?** Die Antwort hängt von vier Faktoren ab: Latenzanforderung, Datenvolumen, Konnektivitätsverfügbarkeit und Datenschutz/Souveränität.

Stärken der Cloud-Verarbeitung	Stärken der Edge-Verarbeitung
✓ Hohe Rechenleistung für komplexe Modelltraining	✓ Latenz < 10 ms — kritisch für Sicherheits- und Regelkreise
✓ Zentrale Datenhaltung für fleetweite Analysen	✓ Offline-Fähigkeit: funktioniert ohne WAN-Verbindung
✓ Skalierung ohne lokale Hardware-Investition	✓ Bandbreitenreduktion: nur relevante Ereignisse werden gesendet
✓ Langzeitarchivierung und Compliance-Reporting	✓ Datensouveränität: Rohdaten verlassen das Werk nicht

✓ Einfaches Modell-Update über Fernzugriff ✓ Ideal für Batch-Analysen und Dashboards	✓ Deterministische Reaktionszeiten für OT-Steuerung ✓ Geringere laufende Cloud-Kosten bei hohem Datenvolumen

1.3 Der Edge-AI-Markt: Zahlen und Wachstum

Der globale Edge-AI-Markt befindet sich in einer Wachstumsphase, die strukturell von der Reifung der Fertigungsautomatisierung, der Verbreitung von 5G-Konnektivität und der Miniaturisierung von KI-Chips getrieben wird:

Kennzahl	Wert / Prognose	Quelle / Stand
Globaler Edge-AI-Markt 2025	ca. 25 Mrd. USD	Mender.io / Branchenanalysen 2026
Globaler Edge-AI-Markt 2033 (Prognose)	ca. 120 Mrd. USD	CAGR ~22 % p.a.
Latenz: Cloud-Roundtrip (typisch)	100–500 ms	Industriepraxis / AWS IoT Benchmarks
Latenz: Edge-Inferenz (lokal)	< 10 ms	NVIDIA Jetson Benchmarks 2025
TinyML: RAM-Bedarf für CNN-Inferenz	< 1 MB (quantisiert)	TinyML Systematic Review, MDPI 2026
Anteil Hersteller mit Edge-AI-Piloten (2025)	ca. 62 %	Gartner IIoT Survey 2025
Predictive-Maintenance-ROI (typisch)	10–40 % Wartungskostenreduktion	AlphaBOLD / McKinsey Industrieanalysen

2. thin-edge.io: Das Open-Source-Framework für industrielle Edge-Agenten

2.1 Was ist thin-edge.io?

thin-edge.io ist ein quelloffenes, modulares Framework für schlanke IoT-Edge-Agenten auf Linux-basierten Geräten. Ursprünglich von Cumulocity (Software AG) entwickelt, wurde es 2022 als Eclipse-Foundation-Projekt eingebracht und wird heute von einer wachsenden Community von Industrieunternehmen, Systemintegratoren und Plattformanbietern weiterentwickelt.

Das Ziel von thin-edge.io ist präzise definiert: Ein Linux-basiertes IoT-Gerät — vom Raspberry Pi bis zum industriellen IPC — soll **sicher, zuverlässig und mit geringem RAM-Footprint** mit jeder gängigen IoT-Plattform (AWS IoT, Azure IoT, Cumulocity und anderen) verbunden werden können. Die v1.0-Version (2024) markiert den Übergang von Experimentier- zu Produktionsreife.

thin-edge.io — Kerneigenschaften

- Implementiert in Rust: native Performance, minimaler Speicherbedarf, sichere Speicherverwaltung
- Multicloud: out-of-the-box Support für AWS IoT Core, Azure IoT Hub, Cumulocity — erweiterbar
- MQTT-basierter Nachrichtenbus: modulare Erweiterungen in beliebiger Programmiersprache andockbar
- OTA-Management: Software-, Firmware- und Konfigurations-Updates over-the-air
- Protokoll-Plugins: OPC-UA, Modbus, Node-RED — industrielle Protokolle out-of-the-box
- Container-fähig: läuft in Docker/Podman auf Embedded Linux und Industrial PCs
- Security by Design: TLS, Zertifikats-Management, Secure Boot Integration

2.2 Architektur: Der MQTT-basierte Nachrichtenbus

Die Architektur von thin-edge.io folgt dem Prinzip des **lose gekoppelten Nachrichtenbusses**. Alle Komponenten — Datenquellen, Protokollübersetzer, KI-Inferenz-Engines, Cloud-Connectoren — kommunizieren über einen lokalen MQTT-Broker. Das macht das System modular, wartbar und in beliebiger Programmiersprache erweiterbar.

Datenquellen	▸ Maschinensteuerungen (PLC, SPS) via OPC-UA oder Modbus ▸ Sensoren: Vibration, Temperatur, Druck, Strom, Bildverarbeitung ▸ Externe Systeme: MES, SCADA, ERP via REST-API ▸ Proprietäre Protokolle via Custom-Plugin (Rust, Python, C)
thin-edge.io Core	▸ Lokaler MQTT-Broker (Mosquitto): interner Nachrichtenbus ▸ Protokoll-Mapper: OPC-UA → MQTT, Modbus → MQTT, Sparkplug B ▸ Telemetry Agent: Datenerfassung, Pufferung, Kompression ▸ Software Management Agent: OTA-Updates, Plugin-Lifecycle ▸ AI Inference Plugin: ONNX Runtime, TensorFlow Lite, Edge Impulse

	▸ Node-RED Integration: Low-Code-Datenverarbeitung und Routing
Cloud / Platform	▸ Cumulocity IoT: Gerätemanagement, Dashboards, Analytics ▸ AWS IoT Core / Azure IoT Hub: Cloud-native Integration ▸ Modell-Deployment: KI-Modelle via OTA auf Edge-Geräte deployen ▸ Fleetweites Monitoring: Aggregation von Telemetrie aus 1000+ Geräten

2.3 Thin Edge vs. Thick Edge: Die richtige Wahl

Die Wahl zwischen Thin-Edge- und Thick-Edge-Ansatz ist keine Ideologiefrage, sondern eine Funktion der Anforderungen. Die folgende Orientierung hilft bei der Entscheidung:

Kriterium	Thin Edge (thin-edge.io)	Thick Edge (NVIDIA Jetson, IPC mit GPU)
Hardware	Raspberry Pi, ARM-Cortex, Industrial Gateway, < 512 MB RAM	NVIDIA Jetson Orin, Intel NUC mit CUDA-GPU, > 4 GB RAM
Inferenz-Komplexität	TinyML, quantisierte Modelle, ONNX-Lite, Anomalieerkennung	Computer Vision, LLMs, komplexe Zeitreihen-DNN, Echtzeit-Video
Latenz	< 10 ms (lokal), deterministisch	< 5 ms (GPU-beschleunigt)
Kosten pro Gerät	50–300 EUR (Gateway/SBC)	800–5.000 EUR (Jetson Orin und höher)
Anwendung	Predictive Maintenance, Anomalieerkennung, Protokollübersetzung	Qualitätssichter via Kamera, Robotik-Steuerung, Sprachsteuerung
Deployment-Skala	100–10.000 Geräte, zentral über OTA verwaltet	Typisch 10–100 Geräte, höhere IT-Betriebsanforderungen

3. KI-Technologien am industriellen Edge

3.1 Das KI-Methodenspektrum für Edge-Deployment

Nicht jede KI-Methode ist am Edge einsetzbar — und nicht jede Aufgabe braucht ein großes Modell. Das folgende Spektrum ordnet Methoden nach Ressourcenbedarf und Anwendungseignung:

Methode	RAM-Bedarf	Typische Aufgabe	Edge-Geeignetheit
Regelbasierte Systeme (Schwellwerte)	< 1 KB	Einfache Alarmlogik, Grenzwertüberwachung	✓✓✓ Überall einsetzbar
TinyML (quantisierte Modelle)	10 KB – 1 MB	Anomalieerkennung, Klassifikation, Vibrations-FFT	✓✓✓ Thin Edge / MCU
ONNX Runtime (optimiert)	10–200 MB	Zeitreihenprognose, Multivariate Anomalie, Regression	✓✓ Industrial Gateway / SBC
TensorFlow Lite / Edge Impulse	1–50 MB	Sensor-Fusion, Audio-Klassifikation, leichte CV	✓✓ Thin/Thick Edge
Computer Vision (YOLO, EfficientDet)	100 MB – 2 GB	Sichtprüfung, Objekterkennung, Defekterkennung	✓ Thick Edge (GPU nötig)
LLM / GenAI (quantisiert)	2–8 GB (GGUF Q4)	Operator-Assistenz, Dokumentensuche, Alarmanalyse	△ Thick Edge, hohe Last
Federated Learning	Variabel	Modelltraining ohne Rohdaten-Transfer	✓ Koordination Edge-Cloud

3.2 TinyML: KI auf Mikrocontrollerebene

TinyML bezeichnet die Ausführung von Machine-Learning-Inferenz auf Mikrocontrollern (MCUs) mit extrem begrenzten Ressourcen — typischerweise < 1 MB Flash, < 256 KB RAM. Durch Quantisierung (INT8/INT4 statt FLOAT32) und Pruning schrumpfen Modelle auf Bruchteile ihrer ursprünglichen Größe, ohne wesentliche Genauigkeitsverluste.

Für Industrieanwendungen bedeutet TinyML: KI direkt im Sensor-Node, ohne lokales Gateway als Zwischenstufe. Ein Vibrationssensor mit TinyML erkennt Lagerdefekte autonom — mit Batterielaufzeiten von Monaten. Die Systematic Review des MDPI-Journals (April 2026, 35 Studien 2018–2026) bestätigt: TinyML ist primär für Predictive Maintenance, Anomalieerkennung und Energiemanagement in IIoT-Umgebungen geeignet.

TinyML-Ökosystem 2026

Frameworks: Edge Impulse (Deployment auf Arduino, Nordic, ST), TensorFlow Lite Micro, ONNX Micro, Microsoft ONNX Runtime

Hardware: Arduino Nano 33 BLE, Nordic nRF52840, STM32, Raspberry Pi Pico, ESP32-S3

Neuronimorphe Chips (2025+): Innateras Pulsar — 26 TOPS bei 2,5 W; Intel Loihi 2

Anwendung: Vibrations-FFT, Akustische Anomalie, Temperaturprognose, Energieverbrauchsklassifikation

3.3 Protokolle als Enabler: OPC-UA trifft MQTT

In der industriellen Praxis treffen zwei grundlegend verschiedene Protokollwelten aufeinander: **OPC-UA** (das semantisch reiche, maschinennahe Protokoll der OT-Welt) und **MQTT** (das schlanke, asynchrone Transportprotokoll der Cloud-Integration). Die moderne Edge-Architektur verbindet beide — mit dem Edge-Gateway als Übersetzer.

Das Paradigma 2026: OPC-UA auf der Maschinenebene (LAN), MQTT als Transport zur Cloud. Der Edge-Agent (thin-edge.io) abonniert OPC-UA-Tags, strippt den Overhead und publiziert als MQTT-Payload. Ergebnis: semantisch strukturierte Daten in der Cloud, ohne proprietäre Protokoll-Abhängigkeit.

4. AIoT-Anwendungsfälle in der industriellen Praxis

Die folgenden fünf Use Cases repräsentieren die reifsten und wirtschaftlich relevantesten AIoT-Anwendungsfelder in der Fertigungs- und Prozessindustrie. Sie sind keine Zukunftsvision, sondern dokumentierte Praxis — mit messbaren Ergebnissen.

Use Case 1 Predictive Maintenance: Lagerüberwachung *[Fertigungsindustrie]*	
Problem / Ausgangslage	Ungeplante Maschinenstillstände durch Lagerdefekte kosten in der Fertigungsindustrie typischerweise 5.000–50.000 EUR/Stunde. Klassische Wartungsintervalle (präventiv, kalenderbasiert) unter- oder überversorgen die Anlage.
Edge-AI-Lösung	TinyML-Modell auf Vibrationssensor (Edge Impulse, ONNX Micro): FFT-Analyse des Vibrationssignals, Erkennung von Lagerdefekt-Mustern (Wälzfrequenzen) direkt auf dem Sensor-Node. Thin-edge.io publiziert nur Anomalieereignisse — kein Rohdaten-Upload. Verbleibende Nutzungsdauer (RUL) wird lokal berechnet.
Messbares Ergebnis	Reduktion ungeplanter Stillstände um 35–55 %, Wartungskostenreduktion um 20–40 %, Batterielaufzeit Sensor-Node > 6 Monate, Cloud-Bandbreitenbedarf –95 % gegenüber Raw-Data-Streaming.

Use Case 2 Qualitätssicherung: Visuelle Sichtprüfung *[Automotive / Elektronik]*	
Problem / Ausgangslage	Manuelle Sichtprüfung in Hochgeschwindigkeits-Fertigungslinien ist fehleranfällig, langsam und personalintensiv. Defektrate von 0,1–1 % führt zu Rückrufkosten und Qualitätsstrafen.
Edge-AI-Lösung	Computer-Vision-Modell (YOLO-basiert, EfficientDet) auf NVIDIA Jetson Orin am Ende der Fertigungslinie. Thick-Edge-GPU verarbeitet Kamerabilder in < 5 ms. Lokale Entscheidung: i.O. / n.i.O. — Teile werden inline ausgeschleust. thin-edge.io synchronisiert Ergebnisse und Modell-Updates in die Cloud.
Messbares Ergebnis	Erkennungsrate > 99,2 % (vs. 94–97 % manuell), Durchsatz +40 % gegenüber manueller Prüfung, 0-Latenz-Ausschleusung ohne Produktionsstopp, ROI-Amortisation typisch 8–14 Monate.

Use Case 3 Energiemanagement: Edge-KI zur Lastoptimierung *[Prozessindustrie / Chemie]*	
Problem / Ausgangslage	Energiekosten machen in der chemischen Industrie 20–40 % der Produktionskosten aus. Spitzenlastzuschläge und unoptimierte Anlagenfahrweisen erzeugen signifikante Zusatzkosten.
Edge-AI-Lösung	ML-Regressionsmodell (ONNX Runtime auf Industrial IPC): Prognose des Energieverbrauchs der nächsten 15–60 Minuten basierend auf Produktionsprogramm, Außentemperatur und historischen Mustern. Edge-Agent steuert Anlagenbetriebsparameter (Temperatur, Druck, Durchfluss) innerhalb definierter Betriebsgrenzen autonom. Thin-edge.io sendet Optimierungsprotokoll in die Cloud.

Messbares Ergebnis	Energiekostensenkung 8–15 %, Vermeidung von Lastspitzenzuschlägen, CO_2-Emissionsreduktion proportional. ROI-Amortisation 12–24 Monate je nach Energiepreislage.

Use Case 4 Anomalieerkennung: Prozessüberwachung in Echtzeit *[Öl & Gas / Wasser / Energie]*	
Problem / Ausgangslage	Prozessabweichungen in kritischen Infrastrukturen (Druckbehälter, Pumpen, Ventile) können Sekunden nach Auftreten kritisch werden. Cloud-basierte Überwachung mit 100–500 ms Latenz ist zu langsam für Schutzabschaltungen.
Edge-AI-Lösung	Multivariate Anomalieerkennung (Autoencoder, ONNX Runtime) auf Edge-Gateway: kontinuierliche Überwachung von 50–200 Sensorkanälen. Abweichung vom Normalzustand wird in < 5 ms erkannt und löst lokale Schutzreaktion aus. Thin-edge.io sendet Anomaliereport mit Kontext (Sensorwerte, Zeitstempel, Schweregrad) in die Cloud für Post-Mortem-Analyse.
Messbares Ergebnis	Reaktionszeit 50–100× schneller als Cloud-basiert, Vermeidung von Sicherheitsvorfällen und regulatorischen Strafen, lückenlose Dokumentation für Behördennachweis.

Use Case 5 KI-gestützter Operator-Assistent (GenAI am Edge) *[Diskrete Fertigung / Maschinenbau]*	
Problem / Ausgangslage	Produktionsmitarbeiter verlieren durchschnittlich 15–30 Minuten pro Schicht mit der Suche nach Informationen (Handbücher, Fehlerprotokolle, Prozessparameter) über heterogene Systeme (ERP, MES, SCADA, PLC).
Edge-AI-Lösung	Quantisiertes LLM (Mistral-7B GGUF Q4, 3,8 GB) auf Raspberry Pi 5 oder Industrial IPC: Natural-Language-Interface für Produktionsmitarbeiter. Abfragen über Sprache oder Text; KI greift via OPC-UA (PLC), MQTT (Sensoren) und REST-API (MES, ERP) auf Echtzeitdaten zu. Latenz KI-Inferenz: ~10 s für komplexe Mehrschritt-Anfragen.
Messbares Ergebnis	Informationszugriffszeit –80 % (15–30 Min. → 2–3 Min.), Fehlerdiagnosezeit –60 %, höhere Mitarbeiterzufriedenheit. Praxisnachweis: Industriepublikation PMC 2025 (Betonwerk Frumecar, Raspberry Pi 5).

5. Architekturleitfaden: Edge-AI-Deployment in der Praxis

5.1 Die vier Entscheidungsdimensionen

Vor dem Start eines Edge-AI-Projekts sind vier Entscheidungen zu treffen, die alle nachfolgenden Technologie- und Investitionsentscheidungen prägen:

Dimension	Leitfrage	Thin Edge	Thick Edge
Latenz	Wie schnell muss entschieden werden?	< 10 ms ausreichend	< 5 ms oder GPU-Inferenz nötig
Modellkomplexität	Welche KI-Methode ist nötig?	TinyML, ONNX Lite, Regelbasiert	CV, DNN, LLM, Echtzeit-Video
Skalierung	Wie viele Geräte werden deployt?	100–10.000, OTA-verwaltet	< 100, höhere Betriebskosten
Datensouveränität	Wo dürfen Daten verarbeitet werden?	Rohdaten bleiben lokal	Rohdaten lokal, Ergebnisse in Cloud

5.2 Der Referenz-Technologie-Stack

Ein industrietauglicher Edge-AI-Stack besteht aus vier Schichten — von der Sensorebene bis zur Cloud-Integration:

Sensor & OT	▸ Sensoren: Vibration (MEMS), Temperatur (PT100/NTC), Strom (Rogowski), Kamera (IDS, Basler) ▸ Steuerungen: Siemens S7, Beckhoff TwinCAT, Mitsubishi MELSEC via OPC-UA ▸ Legacy-Anbindung: Modbus RTU/TCP, PROFINET, EtherNet/IP via Protokoll-Adapter
Edge Gateway	▸ Hardware: Raspberry Pi 5 (Prototyp/KMU), Beckhoff CX, Advantech UNO, NVIDIA Jetson (CV) ▸ OS: Debian/Ubuntu für SBC, Yocto für embedded Linux, Alpine für Container ▸ thin-edge.io: MQTT-Bus, Protokoll-Mapper (OPC-UA/Modbus), OTA-Management ▸ KI-Engine: ONNX Runtime (CPU), TensorFlow Lite, Edge Impulse SDK ▸ Node-RED: Low-Code-Datenrouting, Datenvorverarbeitung, Visualisierung
Edge Analytics	▸ Lokale TSDB: InfluxDB / SQLite für Zeitreihenpufferung und Offline-Betrieb ▸ Inferenz-Engine: ONNX Runtime, TFLite — Modelle via OTA deployt ▸ Ergebnis-Routing: Lokale Aktion (Alarm, Relais) + MQTT-Event an Cloud

	▸ Federated Learning Client: Lokales Modellupdate ohne Rohdaten-Transfer
Cloud & Platform	▸ IoT-Plattform: Cumulocity (thin-edge.io nativ), AWS IoT Core, Azure IoT Hub ▸ Modell-Lifecycle: MLflow, Weights & Biases — Training in Cloud, Deployment via OTA ▸ Dashboards: Grafana, Cumulocity Analytics Builder, Power BI ▸ Langzeitarchivierung: Azure Data Lake, AWS S3 — für Compliance und Retraining

5.3 Sicherheit in der Edge-AI-Architektur

Edge-Deployment erhöht die Angriffsfläche: Tausende physisch zugänglicher Geräte, oft in ungeschützten Umgebungen. Ein Security-by-Design-Ansatz ist von Anfang an einzuplanen:

- Secure Boot: Verhinderung unbefugter Firmware auf Edge-Geräten
- TLS-verschlüsselte MQTT-Verbindung: Alle Cloud-Kommunikation über MQTT mit Zertifikat
- Zero-Trust-Prinzip: Jedes Edge-Gerät authentifiziert sich individuell — kein 'shared secret'
- OTA-Signierung: Software-Updates werden kryptografisch signiert — thin-edge.io prüft vor Installation
- Netzwerksegmentierung: OT-Netz (Maschinen) und IT-Netz (Cloud-Konnektivität) getrennt — DMZ-Konzept
- NIS2 / IEC 62443: Edge-Komponenten fallen unter OT-Cybersicherheitspflichten für KRITIS-Betreiber

6. Wirtschaftliche Bewertung: Von der Idee zum Business Case

6.1 Der dreistufige Bewertungsrahmen

Ein belastbarer Business Case für Edge-AI-Investitionen folgt drei Bewertungsstufen: Potenzialabschätzung → Pilotvalidierung → Skalierungsrechnung. Nur wer alle drei Stufen durchläuft, kann eine fundierte Investitionsentscheidung treffen.

Stufe	Frage	Methode	Ergebnis
1 — Potenzialabschätzung	Lohnt sich Edge-AI für diesen Use Case überhaupt?	ROI-Schnellkalkulation, Benchmark-Vergleich, Technologie-Readiness	Go/No-Go Entscheidung für Pilot
2 — Pilotvalidierung	Wie viel Nutzen erzielt das Modell in unserer Umgebung?	6–12 Wochen Pilot, Baseline vs. Edge-AI, KPI-Messung	Validierter Business Case mit echten Zahlen
3 — Skalierungsrechnung	Was kostet der Rollout — und was bringt er?	TCO-Modell: Hardware + Software + Betrieb + Training vs. Nutzen	Investitionsfreigabe und Rollout-Planung

6.2 Typische ROI-Treiber und Kostenpositionen

Nutzenpositionen (Beispiel)	Kostenpositionen (Beispiel)
▸ Stillstandsvermeidung: 5.000–50.000 EUR/h × vermiedene Stunden	▸ Hardware: 50–5.000 EUR/Gerät je nach Thin/Thick Edge
▸ Wartungskostenreduktion: 20–40 % auf Instandhaltungsbudget	▸ Rollout-Aufwand: Installation, Konfiguration, Integration
▸ Qualitätskosten: Rückruf- und Nacharbeitskosten, Ausschussreduktion	▸ KI-Entwicklung: Modelltraining, Validierung, Dokumentation
▸ Energieeinsparung: 8–15 % auf Energiekostenposition	▸ Cloud-Plattform: SaaS-Lizenz oder selbst gehostete Infrastruktur
▸ Bandbreitenkosten: Reduktion Cloud-Datentransferkosten –80–95 %	▸ OTA-Management: thin-edge.io + Plattform-Lizenz
▸ Arbeitszeitgewinn: Operator-Effizienz, weniger Fehlersuche	▸ Laufender Betrieb: Modell-Retraining, Updates, Support

6.3 Industrie-Benchmarks

Die folgenden Benchmarks basieren auf publizierten Industriestudien und Praxisberichten. Sie sind Orientierungswerte — keine Garantien für den Einzelfall:

Use Case	Typischer ROI-Zeitraum	Einsparpotenzial	Datenquelle
Predictive Maintenance (Lager/Motor)	12–24 Monate	10–40 % Wartungskosten, 35–55 % weniger Stillstände	McKinsey / AlphaBOLD 2026
Visuelle Qualitätssicherung	8–14 Monate	Ausschussreduktion 60–80 %, Prüfzeit –40 %	Embedded Computing Design 2025
Energiemanagement (Lastoptimierung)	12–24 Monate	8–15 % Energiekostenreduktion	Industriepraxis / EnBW-Studien
Anomalieerkennung Prozess	6–18 Monate	Vermeidung 1–3 Sicherheitsvorfälle/Jahr	IEC 61511 Benchmarks
Operator-Assistent (GenAI)	18–36 Monate	–80 % Informationssuchzeit, Schulungskosten	PMC 2025

7. Implementierungsleitfaden: Edge-AI in fünf Schritten

Der folgende Leitfaden beschreibt einen praxiserprobten Implementierungspfad für industrielle Edge-AI-Projekte — von der ersten Use-Case-Identifikation bis zum skalierten Rollout.

Schritt 1 — Use-Case-Identifikation und Priorisierung (2–4 Wochen)

- Workshop mit Produktion, Maintenance und IT: Wo sind die größten ungeplanten Stillstände / Qualitätskosten?
- Daten-Verfügbarkeitscheck: Gibt es historische Sensordaten für Modelltraining? Welche Protokolle sprechen die Maschinen?
- Quick-ROI-Kalkulation: Stunden Stillstand/Jahr × Kosten/h = maximales Nutzenpotenzial
- Technologie-Readiness: Thin Edge oder Thick Edge? Welche Hardware ist bereits vorhanden?

Schritt 2 — Dateninfrastruktur aufbauen (4–8 Wochen)

- thin-edge.io auf Pilot-Gateway installieren (Raspberry Pi 5 oder Industrie-Gateway)
- Maschinenanbindung: OPC-UA-Server konfigurieren oder Modbus-Adapter installieren
- Datenerfassung starten: Baseline-Daten (Normalbetrieb, historische Anomalien) sammeln
- MQTT-Broker testen: Datenfluss vom Sensor zur Cloud validieren

Schritt 3 — KI-Modell entwickeln und validieren (6–12 Wochen)

- Feature-Engineering: FFT, gleitende Mittelwerte, Differenzialgrößen aus Rohdaten
- Modellauswahl: TinyML für MCU, ONNX Runtime für Gateway — immer mit Ressourcen-Budget
- Trainingsumgebung: Cloud-Training mit historischen Daten, Export als ONNX oder TFLite
- Edge-Validierung: Modell auf Zielgerät deployen, Inferenzzeit und RAM-Bedarf messen
- A/B-Vergleich: KI-Entscheidung vs. Expertenurteil über 4–8 Wochen — Precision/Recall messen

Schritt 4 — Pilotbetrieb und Business-Case-Validierung (8–16 Wochen)

- Produktiveinsatz im definierten Pilotbereich — nicht in der gesamten Fertigung
- KPI-Tracking: Vermiedene Stillstunden, Defekterkennungsrate, Energieverbrauch
- OTA-Update-Prozess testen: Modell-Verbesserung im Betrieb deployen
- Stakeholder-Reporting: Monatlicher ROI-Bericht für Entscheider

Schritt 5 — Skalierung und Betrieb (ab Monat 6)

- Rollout-Planung: Geräteanzahl, Rollout-Zeitplan, OTA-Batch-Management

- Modell-Lifecycle: Retraining-Intervalle, Drift-Erkennung, Versions-Management
- Sicherheits-Härtung: Zertifikats-Rotation, Netzwerksegmentierung, Audit-Log
- Governance: Wer ist verantwortlich für Modellqualität? PMO-Integration (Strategic Portfolio Management)

8. Strategische Einordnung: Edge-AI im Unternehmensportfolio

Edge-AI-Projekte sind keine IT-Projekte — sie sind strategische Transformationsinitiativen, die Produktionsprozesse, Geschäftsmodelle und Wettbewerbspositionen verändern. Die strategische Einordnung im Portfolio ist entsprechend wichtig.

8.1 Typische Reifegradentwicklung

Reifegrad	Charakteristika	Typische Organisationen
Stufe 1 — Konnektivität	Maschinen sind vernetzt, Daten werden gesammelt, kein Analytics	Traditionelle Fertigung, KMU ohne IT-Ressourcen
Stufe 2 — Monitoring	Dashboards, Schwellwertalarme, reaktive Wartung	Mittlere Fertigungsbetriebe mit IIoT-Einstieg
Stufe 3 — Analytics	Cloud-basierte Analysen, erste ML-Modelle, Predictive Maintenance Pilot	Industrieunternehmen mit dediziertem OT/IT-Team
Stufe 4 — Edge-AI	KI-Inferenz am Edge, Echtzeit-Entscheidungen, OTA-Modell-Management	Digitalisierungsführer, Industry-4.0-Vorreiter
Stufe 5 — Autonomie	Selbstlernende Systeme, federated Learning, Human-AI-Kollaboration	Industry 5.0 — wenige Vorreiter weltweit

8.2 Erfolgsfaktoren für industrielle Edge-AI

- **OT/IT-Alignment: Edge-AI scheitert häufig nicht an der Technologie, sondern an der Trennung von OT und IT. Gemeinsame Governance ist entscheidend.**
- Datenstrategie vor Technologiestrategie: Schlechte Datenbasis erzeugt schlechte Modelle — egal wie leistungsfähig die Hardware.
- Use-Case-Fokus: Nicht 'KI im gesamten Werk', sondern ein konkreter, messbarer Use Case mit klarem Business Case.
- Edge-Security von Anfang an: Sicherheitsnachrüstung ist teurer als Security by Design.
- Modell-Lifecycle-Management: Ein deployetes Modell altert — Drift-Erkennung und Retraining-Prozesse müssen von Anfang an geplant werden.
- Portfolio-Integration: Edge-AI-Initiativen gehören ins strategische Portfolio — mit Sponsor, Budget und Meilensteinplan (SPM-Logik).

9. Fazit: Edge-AI als Wettbewerbsvorteil

KI am industriellen Edge ist 2026 kein Experiment mehr — sie ist einsatzreif, wirtschaftlich sinnvoll und für viele Industrieunternehmen der nächste logische Schritt auf dem Weg zur wettbewerbsfähigen, resilienten Produktion.

Das thin-edge.io-Framework hat einen entscheidenden Beitrag zur Industrialisierung geleistet: Es macht den Einstieg erschwinglich (< 300 EUR Hardware für Thin-Edge-Szenarien), herstellerneutral (AWS, Azure, Cumulocity) und sicherheitstechnisch fundiert (Rust, TLS, OTA-Signierung). Damit senkt es die Einstiegshürde für KMU erheblich — Edge-AI ist kein Exklusivthema für Konzerne mehr.

Der strategische Imperativ für Entscheidungsträger: Edge-AI-Initiativen nicht als IT-Projekte behandeln, sondern als strategische Investitionen — mit klarem Use Case, validiertem Business Case und Portfolio-Governance. Wer heute anfängt, systematisch Edge-KI-Kompetenz aufzubauen, sichert sich einen Vorsprung, der in zwei bis drei Jahren schwer aufzuholen sein wird.

Drei Empfehlungen für den Einstieg

1. Beginnen Sie mit einem Use Case, nicht mit einer Plattform: Identifizieren Sie Ihren teuersten ungeplanten Stillstand — das ist Ihr erster Edge-AI-Use-Case.
2. thin-edge.io als Einstiegsplattform: Open Source, herstellerneutral, Raspberry Pi als Pilothardware — der Einstieg kostet weniger als eine Woche Engineering-Zeit.
3. Business Case vor dem Rollout: Validieren Sie ROI im Pilot. Erst wenn echte Zahlen vorliegen, skalieren Sie — das spart Budget und schützt vor Enttäuschungen.

Glossar: Wichtige Begriffe

Begriff	Definition
AIoT	Artificial Intelligence of Things: Kombination aus IIoT-Konnektivität und KI-Inferenz
Edge Computing	Verlagerung von Datenverarbeitung vom Rechenzentrum an den Netzwerkrand (nahe der Datenquelle)
thin-edge.io	Open-Source-Framework (Eclipse Foundation) für schlanke IoT-Edge-Agenten auf Linux
TinyML	Machine Learning auf Mikrocontrollern (< 1 MB RAM) durch Quantisierung und Modell-Optimierung
OPC-UA	OPC Unified Architecture: industrieller Kommunikationsstandard für OT/Maschinenebene (IEC 62541)
MQTT	Message Queuing Telemetry Transport: schlankes Pub/Sub-Protokoll für IoT-Kommunikation (ISO/IEC 20922)
ONNX	Open Neural Network Exchange: offenes Format für den Austausch von ML-Modellen zwischen Frameworks
OTA	Over-the-Air: Fernaktualisierung von Software, Firmware und KI-Modellen auf Edge-Geräten
Federated Learning	Verteiltes ML-Training: Modell lernt lokal auf Gerätedaten, ohne Rohdaten zu übertragen
RUL	Remaining Useful Life: voraussichtliche verbleibende Nutzungsdauer einer Komponente
Thick Edge	Leistungsstarkes Edge-Gerät mit GPU/NPU (z. B. NVIDIA Jetson) für rechenintensive KI-Aufgaben
Digital Twin	Digitales Abbild einer physischen Maschine — ermöglicht Simulation und Optimierung ohne Eingriff

Quellen & Weiterführende Dokumente

- thin-edge.io v1.0 Release: medium.com/thin-edge-io — Production ready OSS device agent, 2024
- Mender.io: IoT in 2026 — Edge AI, growing complexity, and the demand for smarter updates, Februar 2026
- MDPI Sensors: TinyML in Industrial IoT — Systematic Review (35 Studien 2018–2026), April 2026
- AlphaBOLD: AI-Powered Predictive Maintenance in Manufacturing, April 2026
- N-iX: Key Edge AI Trends transforming enterprise tech in 2026, Februar 2026
- PMC / MDPI: Multimodal Cognitive Architecture with Local Generative AI for Industrial Control (Raspberry Pi 5), Dezember 2025
- PMC: AIoT for Next-Generation Predictive Maintenance — Survey, Industry 5.0, 2025
- IndustryX.ai: MQTT vs OPC-UA — The Honest Architecture Guide, Dezember 2025
- Gartner: Cumulocity Platform Reviews & Industrial IoT Magic Quadrant, 2024–2026
- IJOER: Edge Computing and Real-Time Control for IoT — 2026 Scenario, 2026

Über den Autor

Ralf Platvoet ist Diplom-Ökonom und Inhaber von PPI – Platvoet Performance Intelligence. Er berät Organisationen zu Industrial IoT, AIoT-Architektur, Strategic Portfolio Management und Cyber-Compliance. Sein IIoT-Beratungsschwerpunkt liegt auf der wirtschaftlichen Bewertung von Edge-AI-Investitionen und der Governance von IIoT-Transformationsprojekten — mit Branchenerfahrung in Fertigungsindustrie, Energiewirtschaft und Maschinenbau.

Weitere Whitepapers, interaktive Tools und Ressourcen: **platvoet.org**