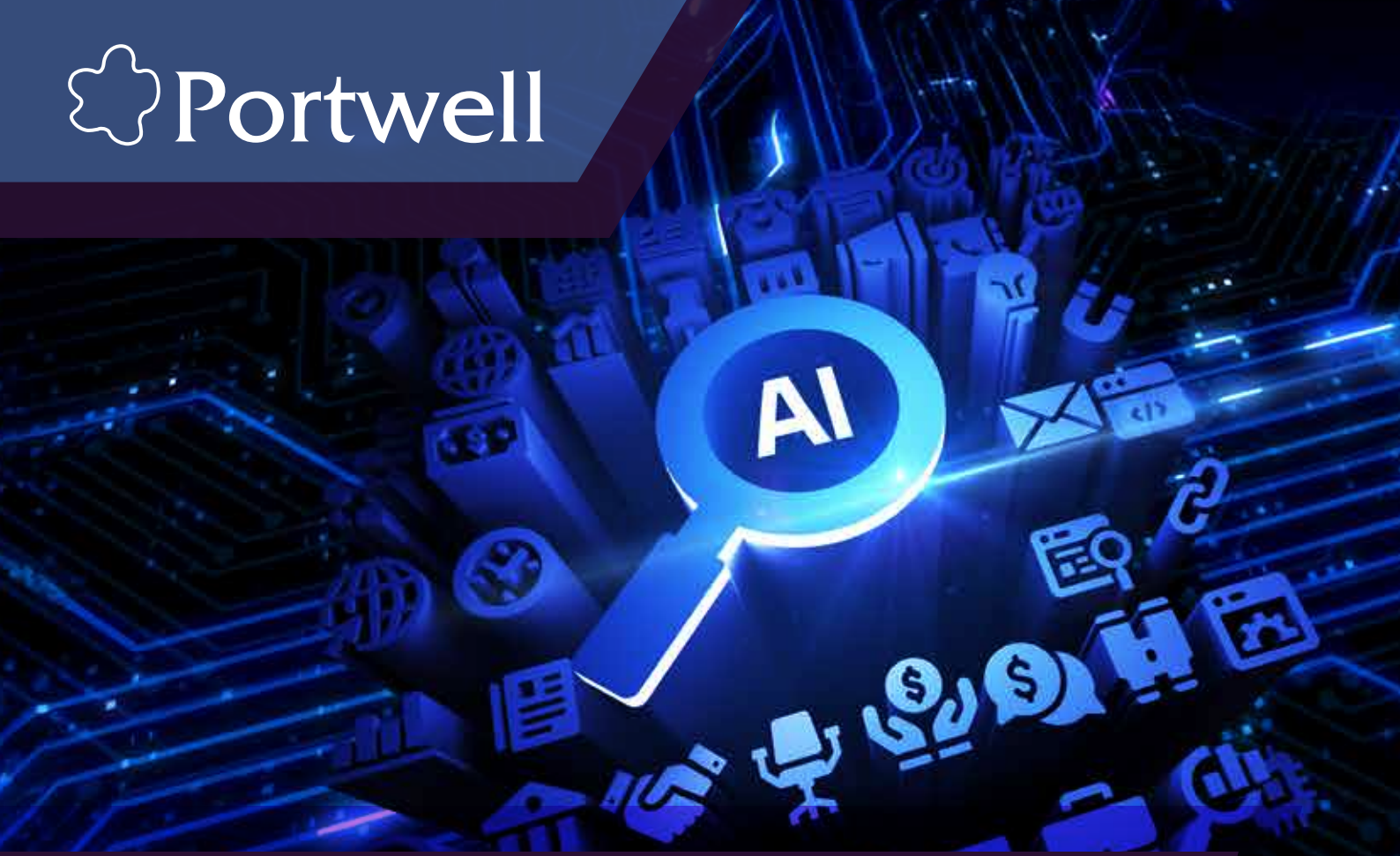
Portwell



From AI Models to Real-World Deployment at the Edge

Building a Scalable Edge AI System Architecture

Presented by
Market Intelligence

2026

Table of Contents

1. AI Is Accelerating Toward Real-World Applications

2. Key Challenges in Deploying AI in Real-World

   2.1 From AI Models to Systems: Core Challenges in AI Deployment

3. Core Edge Computing Capabilities for Real-World AI Deployment

4. Use Cases

   4.1 Low-Latency Decision-Making Deployment :
   How Edge AI Reshapes Operational Value in Smart Fitness Centers

   4.2 Equipment Integration and Edge Computing Collaboration :
   Automated Plastic Recycling and Sorting

5. Portwell DMS: End-to-End Value-Added Services for Accelerating Customer Success

From AI Models to Real-World Deployment at the Edge

1. AI Is Accelerating Toward Real-World Applications

Over the past decade, Artificial Intelligence (AI) has mainly centered on cloud computing, data analytics, and large-scale model training, with applications focused on digital services such as search, recommendations, and business intelligence. In this earlier phase, AI was mainly defined by centralized, cloud-based offline analysis. As sensing, edge computing, high-speed communication, and embedded systems mature, AI is entering a new stage of transformation, moving from virtual environments into real-world settings such as factories, transportation, retail, and unmanned operations. Generative AI, edge computing, and intelligent sensing are accelerating AI adoption across industries, transforming AI from a data analysis tool into a core system for perception, decision-making, and real-time control. Enterprise AI adoption is advancing beyond experiments, enabling real-world operations, on-site optimization, and real-time collaboration with equipment, robots, sensors, and control systems.

In the past, when enterprises evaluated AI, they often focused on model accuracy, parameter scale, and inference speed. However, in real-world deployment environments, successful implementation ultimately depends on the system' s ability to maintain stable, long-term operation under complex and dynamic conditions. The global edge AI market is expected to maintain strong growth over the next decade. This trend is driven by technological advancement, digital transformation, 5G adoption, and the rapid growth of intelligent devices. In scenarios requiring fast response and high stability, enterprises increasingly recognize that critical decisions must be made at the edge rather than relying solely on the cloud. As a result, scalable distributed edge architectures are becoming a key source of future competitiveness.

Despite its promising outlook, deploying edge AI in real-world environments remains challenging for developers and system integrators. The following section examines key challenges across environmental conditions, system integration, lifecycle management, and operations maintenance.

2. Key Challenges in Deploying AI in Real-World

2.1 From AI Models to Systems: Core Challenges in AI Deployment

2.1.1 Edge AI Resilience and Sustained Performance

Edge computing platforms may be required to operate continuously in harsh real-world environments, facing temperature fluctuations, vibration, dust, and electromagnetic interference (EMI). As workloads continue to increase, the thermal stress generated by CPUs, GPUs, or NPUs operating under sustained high loads has become a critical bottleneck for modern edge platforms. Without sufficient thermal management, power

integrity, and mechanical design, systems may experience thermal throttling, resulting in unstable FPS, increased inference latency, or even system abnormalities. These issues can reduce system stability and AI inference performance. Future edge platforms must maintain stable, reliable performance under sustained workloads.

2.1.2 Device Integration, Deterministic Control, and Functional Safety

As edge computing evolves from standalone computing nodes to distributed intelligent systems, edge platforms must connect with a wide range of field devices, including cameras, sensors, programmable logic controllers (PLC), robotic arms, conveyor systems, and human-machine interfaces (HMI). They must also convert inference results into executable control commands. AI models rely on data-driven and statistical inference, so their outputs may vary with data quality, environmental changes, and confidence scores, unlike the fixed rule-based logic of traditional control systems. Therefore, in scenarios involving industrial control, equipment coordination, and functional safety, it is critical to ensure that inference results do not directly drive equipment without proper validation. The system must also meet field-level requirements for deterministic timing, low latency, and low jitter.

2.1.3 Gap Between Long Lifecycles and Rapid Platform Iteration

Industrial equipment and infrastructure typically have lifecycles of more than ten years, while AI SoCs, GPUs, and CPUs are often refreshed within only a few years. This gap between industrial lifecycle requirements and semiconductor upgrade cycles is becoming a major challenge for long-term edge AI deployment. When existing platforms reach End of Life (EOL), enterprises often face significant costs related to Board Support Package (BSP) compatibility, driver updates, AI framework migration, and system revalidation. Future edge platforms must offer high performance, long-term availability, modular expansion, and upgradability to lower redeployment and maintenance costs.

2.1.4 Large-Scale Deployment Operations and System Management

As AI deployment expands from standalone devices to multi-site and global rollouts, the real challenge for enterprises shifts from deployment to maintenance. With large numbers of edge computing nodes distributed across factories, transportation environments, and unmanned sites, any system failure, firmware corruption, or update error can quickly drive up on-site maintenance costs.

3. Core Edge Computing Capabilities for Real-World AI Deployment

Edge computing systems may be required to operate in harsh real-world conditions, including temperature extremes, vibration, dust, and EMI. Therefore, robust thermal design is essential to prevent thermal throttling and performance degradation under high workloads. EMC and EMI testing help ensure signal stability and reliable operation in complex electrical environments. Ruggedized industrial design improves resistance to harsh conditions and long-term wear, reducing downtime and maintenance costs while ensuring system design prioritizes stability and durability beyond computing performance. Since the lifecycle of industrial equipment is typically much longer than the refresh cycle of AI SoCs and GPUs, edge platforms must also support modularity and upgradability to reduce redeployment and revalidation costs.

As AI becomes increasingly integrated with equipment control and OT systems, system-level challenges extend beyond computing performance to include timing, synchronization, reliability, and safety. AI inference results should not directly trigger equipment actions; instead, they should be processed and validated through rule-based control logic, safety mechanisms, and real-time control layers to support predictable, stable, and safe industrial operation. In addition to inference acceleration, the system may also need to support real-time networking, time synchronization, industrial data exchange, and real-time OS capabilities, including TSN, PTP, OPC UA PubSub, and PREEMPT_RT, depending on the control architecture and application requirements. When confidence scores, data quality, or safety conditions are insufficient, the system should enter a conservative mode, trigger re-identification, request human review, or initiate a safe shutdown to support control stability and safety-oriented operation.

Traditional centralized cloud architectures are increasingly unable to meet real-world demands for low latency and real-time response, especially in millisecond-level decision-making scenarios. Transmitting all data to the cloud for inference can introduce network latency, bandwidth constraints, and connection stability risks. As a result, AI deployment is moving toward distributed edge computing, where inference is performed closer to the data source. In this architecture, edge nodes handle real-time inference, anomaly response, and equipment control, while the cloud manages model training, historical analysis, and cross-site optimization. This reduces the amount of data sent back to the cloud and improves system continuity in unstable or high-load network environments

To support large-scale edge deployment, platforms can no longer rely solely on OTA updates. They require remote monitoring, remote recovery, and out-of-band (OOB) management to track device status, temperature, power, network connectivity, and workloads in real time. OOB management enables remote recovery from crashes, OS corruption, or update failures, reducing downtime and on-site maintenance. Future edge platforms will compete on inference performance, observability, maintainability, and operational sustainability.

4. Use Cases

Deployment requirements vary across real-world applications, but all rely on edge computing, real-time inference, system integration, and long-term operations. Therefore, this chapter demonstrates how practical applications leverage real-time inference, system integration, and long-term operations to achieve scalable and sustainable deployment.

4.1 Low-Latency Decision-Making Deployment : How Edge AI Reshapes Operational Value in Smart Fitness Centers

In recent years, competition in the fitness industry has evolved from facility utilization and equipment quantity toward operational efficiency and service consistency. As chain expansion, community gyms, 24/7 operations, and unmanned nighttime services become more common, Operators face inconsistent service quality, rising safety risks, and higher multi-site maintenance costs. Against this backdrop, edge AI is well suited for 24/7 gyms by bringing decision-making and feedback closer to cameras and fitness equipment, rather than relying solely on centralized processing. This enables lower latency and better privacy protection while supporting two high-value use cases: AI coaching for real-time posture feedback and training assistance, and safety monitoring during equipment use with event alerts. When these use cases become systemized, measurable, and scalable across multiple locations, gyms can achieve truly scalable, low-labor, and uninterrupted operations.

Smart Gym Management Architecture： Distributed AI Inference with Centralized Operations

This project follows a typical distributed AI inference and centralized operations management architecture. Multiple on-site video streams are processed by edge AI systems for pose estimation, motion tracking, state detection, and event classification. Only results, not raw video, are sent to the management side for alerts, logs, and multi-site analysis. This reduces cloud streaming cost and latency while supporting privacy protection through edge-based de-identification and selective data upload.
Portwell PJAI edge AI inference systems offer tiered computing performance for gyms of different scales:

A. Larger Gyms: Multi-Channel Video, PoE Power, and High-Bandwidth Uplink

For larger gyms, the PJAI-200 is designed for multi-camera and multi-sensor integration. Powered by the NVIDIA Jetson AGX Orin, it delivers up to 200 TOPS of AI performance and features 32GB LPDDR5 memory, along with rich I/O including 12 PoE GbE ports, 10GbE, and 8 USB 3.2 Type-A ports. For gym operators, this means a single edge AI system can connect video streams from multiple areas, such as free-weight zones, functional training areas, and cardio equipment zones, while PoE simplifies cabling and power management. This is especially valuable for rapid store expansion or retrofitting existing facilities. The upcoming PJAI-2200 will offer up to 275 TOPS of AI performance with rich I/O support, making it suitable for edge AI applications that require higher computing power and multi-channel video transmission.

B. Smaller Gyms: Area-Based Inference Close to Specific Deployment Zones

For smaller gyms or area-based deployments, the PJAI-1100F and PJAI-1100 provide more compact options. The PJAI-1100F, powered by NVIDIA Jetson Orin NX Super Mode, delivers up to 157 TOPS of AI performance. It provides industrial-grade I/O including 2 GbE ports, USB 3.2 Type-A/Type-C, HDMI® 2.0b, 2 COM ports supporting RS-232/422/485, CANbus FD,

8-bit DIO, M.2 expansion, and 12–24V terminal block power input. This makes it suitable for real-time inference and control integration near fitness equipment. For environments that require silent operation, low maintenance, or deployment in areas with dust and fibers, the PJAI-1100 offers a fanless design with a similar I/O configuration and wide-temperature support. It can be installed in low-voltage cabinets, equipment enclosures, or hidden wall-mounted locations.

The upcoming PJAI-1200 extends the PJAI-1100 series with optional PoE, GMSL, and OOB expansion, enhancing flexibility for multi-channel video input, area-level AI analysis, and remote management.

Welcome to Portwell’ s booth during COMPUTEX 2026 to explore the features and application potential of the PJAI series.

OOB Remote Management Solution: Maximizing the Value of Edge AI Device Management and Operations

OOB technology enables remote recovery, troubleshooting, OTA updates, configuration management, and environmental sensing through a unified portal. By integrating Portwell’ s edge AI hardware with OOB remote management, gyms can maintain uptime across unmanned or multi-site locations while reducing on-site maintenance and improving operational efficiency.

4.2 Equipment Integration and Edge Computing Collaboration: Automated Plastic Recycling and Sorting

As environmental awareness rises, plastic recycling has become a key driver of the green economy. Automated sorting systems help reduce reliance on manual labor, improve material identification accuracy, and increase the value of recycled plastics. Accurate classification is especially critical for materials such as PET, HDPE, and PP, while multilayer materials used in textiles, food packaging, and cosmetic packaging add further sorting complexity. By combining AI image recognition, optical sorting, and automation, modern recycling systems can identify complex plastic materials more accurately and improve overall efficiency. Technologies such as color sensors, NIR spectroscopy, and X-ray transmission (XRT) can rapidly distinguish different plastic types, including PET and PS. NIR identifies materials by analyzing the unique spectral signatures created by different chemical bond responses, while color detection helps separate transparent, white, dark, and mixed-color plastics. Transparent and white plastics typically have higher market value, driven by their broader reuse potential and easier dyeing. As a result, accurate material and color classification is essential to improving recycled plastic quality and market value. In automated textile and plastic recycling sorting systems, the edge computing platform must not only perform AI inference, but also integrate NIR cameras, color sensors, lighting modules, conveyor systems, robotic arms, and HMIs.

WEBS-91J0 Enables Multi-Device Integration and Edge AI Collaborative Control for Recycling Sorting Systems

Portwell’ s newly launched WEBS-91J0 embedded system adopts Intel Atom® x7433RE platform, combining low-power computing with a compact, fanless, and ruggedized design. It can serve as the edge computing and control core for textile recycling, material sorting, and automated classification equipment. With a compact size of 152 × 112 × 48 mm and a weight of approximately 1.0 kg, the WEBS-91J0 helps save space inside control cabinets and equipment enclosures. It is well suited for space-constrained industrial environments near conveyor systems, robotic arms, and sorting stations.

To support multi-device coordination in recycling sorting applications, the WEBS-91J0 provides rich I/O, including 3 × RJ45 GbE, 3 × USB 3.0, 1 × RS-232, 1 × RS-485, HDMI 1.4 and DP 1.2. These interfaces enable connection with NIR cameras, color recognition sensors, lighting control modules, HMI displays, barcode readers, conveyor controllers, and other field devices. Through multiple network, USB, and serial communication interfaces, the system can perform image acquisition, sensor data collection, equipment status monitoring, and control command transmission at the edge, helping recycling operators build more flexible automated sorting workflows.

For AI expansion, the WEBS-91J0 features 1 × full-size/half-size mini-PCIe slot and can be configured with a Hailo-8 module to enhance edge AI inference capability. This enables on-site image recognition, material classification, object detection, and defect inspection. With M.2 B-Key 2242/3042/3052 support, the system can be configured with a SATA SSD,

NVMe SSD, or 5G module for local data storage, model deployment, classification record keeping, or wireless data transmission. It also supports 1 × SIM card holder, improving deployment flexibility for remote device management, sorting data upload, and cross-site monitoring. The Intel Atom® x7433RE SKU supports -40°C to 60°C operation, 1.5Grms / 5–500Hz vibration tolerance, and 20G, 11 ms shock tolerance, making it suitable for conveyor equipment, sorting lines, and harsh industrial environments.

The WEBS series combines industrial-grade fanless design, wide-temperature operation, vibration resistance, modular expansion, and rich I/O, with support for 12V or 12–24V DC input and wall-mount or DIN-rail installation depending on the model. Designed for high-dust, high-vibration, and continuous-operation environments, WEBS systems enable edge AI recognition, real-time decision-making, and device integration across recycling sorting, automation, machine vision, smart manufacturing, and equipment monitoring applications, improving efficiency, reliability, and long-term maintenance flexibility.

Welcome to Portwell’ s booth during COMPUTEX 2026 to explore the features and application potential of the WEBS series.

5. Portwell DMS: End-to-End Value-Added Services for Accelerating Customer Success

As AI applications accelerate across real-world environments ,the key question is how to ensure that AI applications can operate stably on site, respond in real time, connect with field devices, support collaborative control, and enable long-term maintenance. To address these complex deployment requirements, enterprises need more than a standalone hardware platform. They need a system integration partner that can help bring AI from proof of concept to mass deployment. With years of DMS (Design & Manufacturing Services) expertise, Portwell provides end-to-end services covering requirement analysis, technical architecture design, prototyping, hardware design, firmware and driver integration, system testing, certification, mass production, and shipment. This helps customers shorten development cycles and reduce deployment risks. Portwell’ s R&D team is experienced in multiple platform architectures and capable of integrating AI accelerator cards, multi-camera inputs, rich I/O, industrial communication, and modular expansion. These capabilities support real-time inference, data acquisition, and device connectivity in real-world environments. In software and firmware integration, Portwell can also support MCU control applications, FPGA-based custom logic, OpenBMC remote management architecture, driver integration, and real-time image recognition applications. This enables complete hardware-software coordination from low-level control to application-layer deployment. Through these capabilities, Portwell helps customers build edge computing systems designed for long-term stable operation, enabling AI to go beyond inference and create sustainable value in real-world operations.

Portwell, Inc. Headquarters

No. 242, Bo'Ai St., Shu-Lin Dist,
New Taipei City 238, Taiwan
Tel: +886-2-7731-8888
Fax: +886-2-7731-9888
E-mail: info@portwell.com.tw
www.portwell.com.tw

Americas

American Portwell (Fremont, CA)

44200 Christy St.,Fremont, CA 94538, USA
Tel: +1-510-403-3399
Fax: +1-510-403-3184
E-mail: info@portwell.com
www.portwell.com

Japan

Portwell Japan, Inc. (Tokyo)

〒112-0011 4-27-10, Sengoku, Bunkyo-ku,
Tokyo, Japan
Tel: +81-3-6902-9225
Fax: +81-3-6902-9226
E-mail: info@portwell.co.jp
www.portwell.co.jp

Portwell Japan, Inc. (Osaka)

〒532-0004 Ste.501 Nippo Shin-osaka Dai-2
Bldg, 1-8-33 Nishi-Miyahara, Yodogawa-ku
Osaka Japan
Tel: +81-6-4807-7721
Fax: +81-6-4807-7720
E-mail: info@portwell.co.jp
www.portwell.co.jp

Europe

European Portwell

Schillingweg 3, 2153 PL
Nieuw Vennep,The Netherlands
Tel: +31-252-620790
E-mail: info@portwell.eu
www.portwell.eu

KIOSK Embedded Systems GmbH

Am Technologiepark 8-10,
D-82229 Seefeld
Tel: +49-8152-3962-500
E-mail: info@portwell.eu
www.portwell.de

Portwell UK Ltd.

Office TH2
Trident House, Trident Park Basil
Hill Road, Didcot, OX11 7HJ, UK
Tel: +44(0)1235-750-760
Fax: +44(0)1235-750-761
E-mail: info@portwell.eu
www.portwell.eu

China

Shanghai Portwell

(201612), Room 1303-1, Building 33, No.258,
Xinhuan Highway, Songjiang District,
Shanghai
Tel: +86-21-5771-2505
Fax: +86-21-5772-2965
E-mail: info@portwell.com.cn
www.portwell.com.cn

Korea

Portwell Korea, Inc.

O-BIZ Tower 1901,No.126,
Beolmal-ro,Dongan-gu,Anyang-si,
Gyeonggi-do, Korea, 431-060
Tel :+82-31-450-3043
Fax:+82-31-450-3044
E-mail: info@portwell.co.kr
www.portwell.co.kr

India

Portwell India Technology Private Limited.

7/3, 3rd Floor, Jeet Dynasty, Swamy
Vivekananda Rd, Nagavarapalya, Bengaluru,
Karnataka 560093
Tel: +91-80-4078-2700
E-mail: enquiry@portwell.in
www.portwell.in

All brand names, product names, and logos may be trademarks or registered trademarks of their respective companies or mark holders.

www.portwell.com.tw