Bringing AI to the Edge: How ML is Powering the Future of IoT

By Gayathri Jegan Mohan
Software Engineer at Microsoft | Azure IoT

What is AI at the Edge mean?

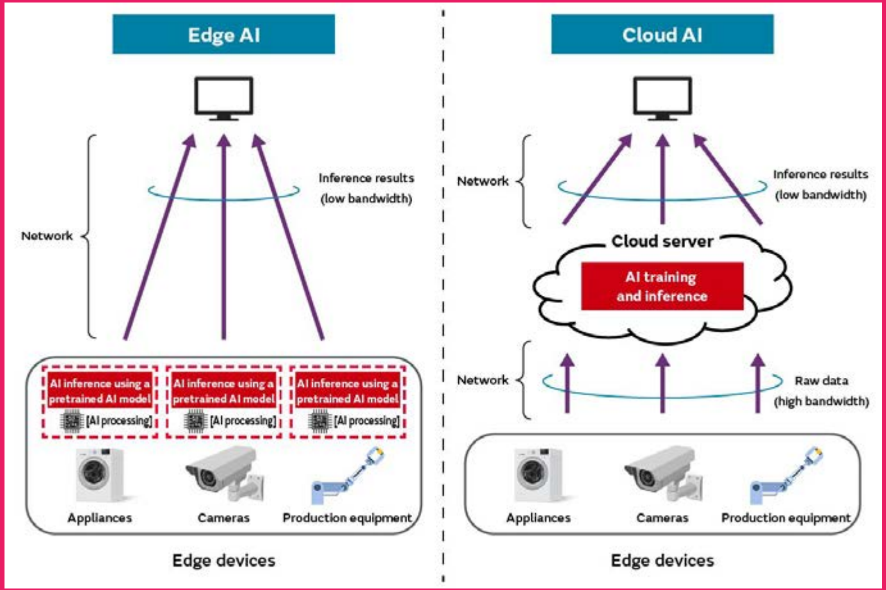
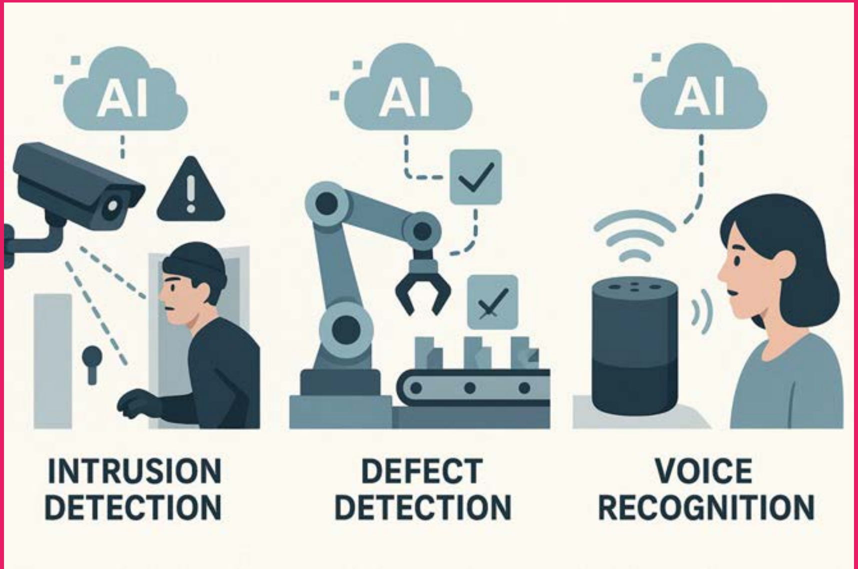


Image @murata.com

- In IoT, usually the ML models are used deployed and managed in cloud.
- But with Edge AI, one can deploy models locally on the edge devices and reduce latency.
- Local inferencing helps in real time processing
- Also solve privacy issues of moving data to cloud.

Examples of Edge AI



- A **surveillance camera** using AI to detect intrusions or license plates on-device.
- A **manufacturing robot** detecting defects in real-time without sending every image to the cloud.
- A **smart speaker** recognizing voice commands locally for faster response and privacy.

Why Edge AI is better than cloud AI?

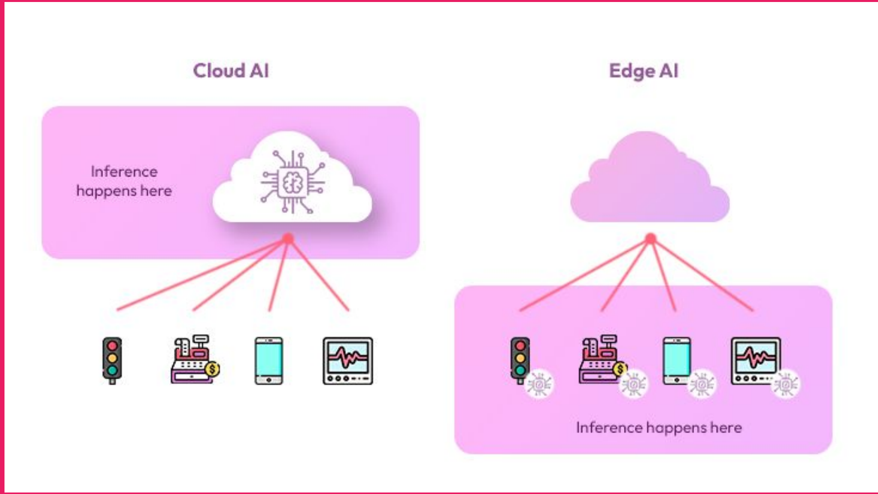
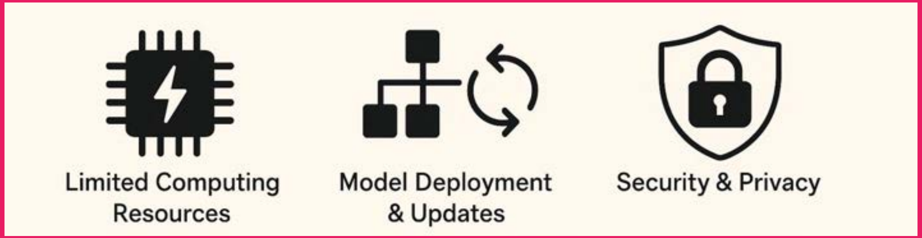


Image @ xailient

- **Latency** - real time response without sending data to cloud
- **Bandwidth** - minimal data sent over network
- **Security** - Data processed locally without cloud risks
- **Reliability** - operates even with limited or no network
- **Energy** - reduces energy usage

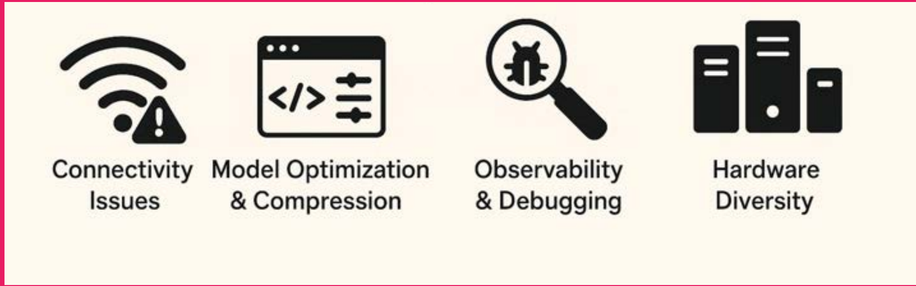
By 2025, 75% of enterprise data will be processed at the edge, revolutionizing AI-powered IoT.

Challenges of Edge AI -(1)



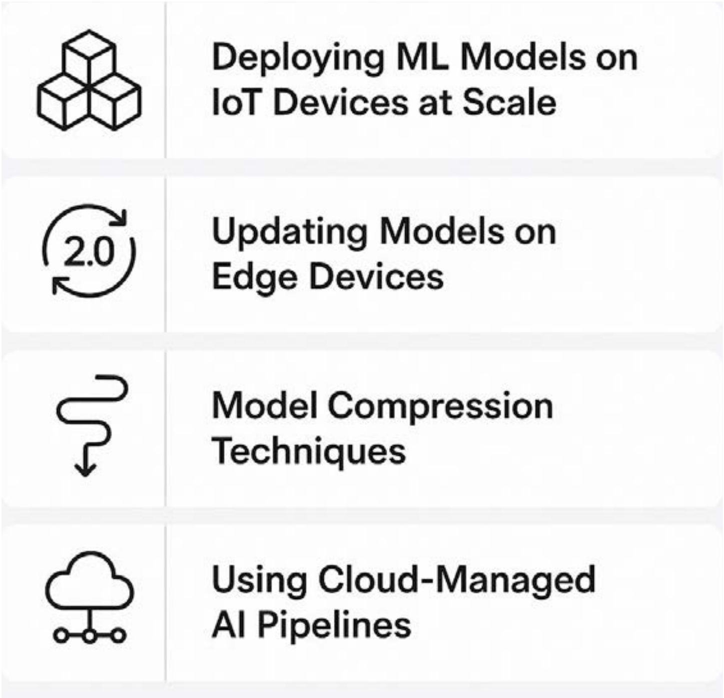
- **Limited Computing Resources**
  - Low processing power compared to GPUs/TPUs
  - Limited memory and storage
  - Power supply (battery operated)
  - Complex models cannot run
- **Model Deployment and updates**
  - Hard to push updates to so many edge devices at onces
  - Risk of version mismatch
- **Security and Privacy**
  - Devices are deployed in public environment prone to malware attacks
  - Need secure boot, secure protocol standards

Challenges of Edge AI -(2)



- **Connectivity Issues**
  - No network sometimes
  - Sync to cloud is hard
- **Model optimization and compression**
  - Models have to be lightweight
  - Accuracy may be comprised with shrinking models
- **Observability**
  - Hard to monitor performance, failures
  - Need simple loggings or simple tools to send data to cloud
- **Hardware Diversity**
  - So many different devices from different vendors
  - Different OS/ runtime

Best Practices of ML models at Edge

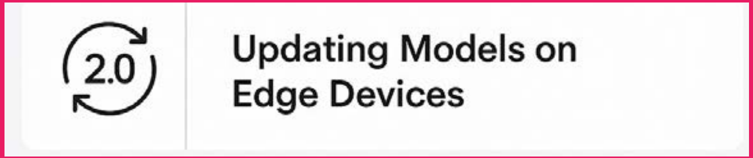


Practice #1 Deploying Models At Edge

When deploying models across thousands of edge devices, it's critical to ensure:

- **Model portability**: Use formats like ONNX or TFLite that work across different hardware.
- **Hardware abstraction**: Target accelerators (e.g., NVIDIA Jetson, Intel Movidius, ARM NPUs) using unified runtimes like OpenVINO or TensorRT.
- **Containerization**: Package models with inference runtimes in Docker containers to ensure consistent execution.

Practice 2# Updating Models on Edge Devices



Frequent updates are required to Improve accuracy, Patch vulnerabilities, Adapt to environmental drift

Some best practices are

- OTA (Over-the-Air) updates with version control
- A/B testing or shadow deployment to evaluate new models without full rollout
- Digital twin simulation to pre-test updates in a cloud replica of your edge environment

Practice 3#
Model Compression Techniques

- **Quantization**
  - Reduces model precision (e.g., from 32-bit float to 8-bit integer)
  - Smaller size model so faster inference
- **Pruning**
  - Removes insignificant weights or neurons from the model
  - Speeds up computation
- **Knowledge Distillation**
  - A small “student” model learns to mimic larger “teacher” model
  - High efficiency with good accuracy

Practice #4
Cloud Managed AI pipelines

Rather than manually managing model lifecycles, modern systems use **cloud-based ML Ops (Machine Learning Operations)** pipelines. These provide:

- **Training and retraining** on the cloud with updated datasets
- **CI/CD pipelines** to automate testing and deployment
- **Telemetry collection** from edge to continuously improve models
- Examples: Azure ML with IoT integration, AWS SageMaker Edge Manager, Google Vertex AI + Edge TPU

Deploying Example

Walmart for deploying models at Scale

- **What they did**: Walmart deployed thousands of cameras with AI models in their retail stores to monitor inventory levels, shelf placement, and customer behavior.
- **How**: Used compact edge servers (NVIDIA Jetson-based) and computer vision models optimized via TensorRT.
- **Result**: Reduced stockouts and optimized restocking schedules, improving operational efficiency across hundreds of locations.

Update Models Example

Tesla Over-the-Air (OTA) Model Updates

What they did: Tesla routinely pushes AI updates (e.g., Autopilot vision and driving behavior models) to cars globally.

How: Uses secure OTA pipelines to validate and deploy updates with rollback capabilities.

Result: Enables incremental model improvements without requiring service center visits, and supports shadow mode testing before full rollout.

Model Compression

Google – MobileNet on Android Devices

What they did: Google developed the MobileNet family—lightweight models designed for mobile inference like image classification or object detection.

How: Used quantization and pruning to reduce model size while retaining accuracy.

Result: Enabled real-time AI features like Google Lens and real-time translation on phones without internet dependency.

Cloud Managed AI Pipeline Example

Siemens Predictive Maintenance in Factories

SIEMENS

What they did: Siemens deployed AI across factory floors for predictive maintenance on CNC machines and motors.

How: Used Azure IoT + Azure ML pipelines to train models in the cloud, then deploy inference versions to edge gateways.

Result: Reduced machine downtime by up to 30%, while using the cloud to continuously retrain models with edge-collected telemetry.

Thank you!