

TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO
SINH HOẠT HỌC THUẬT
TỔNG QUAN VỀ EDGE AI VÀ ỨNG DỤNG TRONG IoT

Người báo cáo: GVC.TS. Nguyễn Duy Huy
Đơn vị: Bộ môn Khoa học máy tính

HÀ NỘI – 5/2026

MỤC LỤC

MỤC LỤC	1
MỞ ĐẦU	7
1. Đặt vấn đề	7
1.2. Xu hướng Edge AI trong hệ thống IoT	7
1.3. Vai trò của Edge AI trong IoT hiện đại	8
1.4. Ứng dụng thực tiễn của Edge AI trong IoT	8
1.4.1. Thành phố thông minh	8
1.4.2. Công nghiệp thông minh	8
1.4.3. Giám sát môi trường	9
1.4.4. Y tế thông minh	9
1.6. Mục tiêu của báo cáo	9
1.7. Phương pháp nghiên cứu	9
CHƯƠNG I: TỔNG QUAN VỀ EDGE AI	11
I.1. Khái niệm về Edge AI	11
I.1.1. Khái niệm Edge Computing	11
I.1.2. Khái niệm Artificial Intelligence (AI)	12
I.1.3. Khái niệm Edge AI	12
I.2. Kiến trúc tổng quát của Edge AI trong IoT	13
I.2.1. Mô hình kiến trúc tổng quát	13
I.2.2. Các thành phần chính	13
I.3. Nguyên lý hoạt động của Edge AI	14
I.4. So sánh Cloud AI và Edge AI	14
I.5. Các nền tảng Edge AI phổ biến	15
I.5.1. Raspberry Pi	15
I.5.2. NVIDIA Jetson	15
I.5.3. Google Coral TPU	16
I.6. Các mô hình AI thường dùng tại Edge	16
I.6.1. CNN (Convolutional Neural Network)	16
I.6.2. YOLO (You Only Look Once)	16
I.6.3. TinyML	16
I.7. Ưu điểm của Edge AI	17
I.7.1. Độ trễ thấp	17

I.7.2. Giảm tải băng thông	17
I.8.1. Hạn chế tài nguyên phần cứng	17
I.8.2. Tiêu thụ năng lượng	18
I.8.3. Khó cập nhật mô hình	18
I.9. Xu hướng phát triển Edge AI	18
CHƯƠNG II: CÁC CÔNG NGHỆ NỀN TẢNG TRONG EDGE AI	19
II.1. Tổng quan về công nghệ nền tảng Edge AI.....	19
II.2. Phần cứng trong Edge AI.....	19
II.2.1. Vai trò của phần cứng Edge AI.....	19
II.2.2. Raspberry Pi.....	19
II.2.3. NVIDIA Jetson	20
II.2.4. Google Coral TPU	21
II.2.5. ESP32 AI và TinyML	22
II.3. Bộ tăng tốc AI (AI Accelerator)	22
II.3.1. Khái niệm.....	22
II.3.2. GPU (Graphics Processing Unit).....	22
II.3.3. TPU (Tensor Processing Unit).....	22
II.3.4. NPU (Neural Processing Unit).....	22
II.4. Framework và phần mềm Edge AI	23
II.4.1. TensorFlow Lite.....	23
II.4.2. PyTorch Mobile	24
II.4.3. OpenVINO	24
II.4.4. Edge Impulse	24
II.5. Các mô hình AI sử dụng tại Edge.....	24
II.5.1. CNN (Convolutional Neural Network).....	24
II.5.2. YOLO.....	24
II.5.3. TinyML	25
II.6. Công nghệ truyền thông trong Edge AI.....	25
II.6.1. MQTT	25
II.6.2. CoAP	25
II.6.3. 5G và 6G	25
II.7. Các kỹ thuật tối ưu mô hình AI	26
II.7.1. Quantization	26

II.7.2. Pruning	26
II.7.3. Knowledge Distillation	26
II.7.4. Model Compression	26
II.8. Bảo mật trong nền tảng Edge AI.....	26
II.9. Xu hướng công nghệ Edge AI	26
CHƯƠNG III: EDGE AI TRONG HỆ THỐNG IoT	27
III.1. Tổng quan về tích hợp Edge AI và IoT	27
III.2. Kiến trúc tích hợp Edge AI trong IoT	27
III.2.1. Kiến trúc ba lớp (Three-layer Architecture).....	27
III.2.2. Device Layer.....	28
III.2.3. Edge Layer.....	28
III.2.4. Cloud Layer	29
III.3. Quy trình hoạt động của hệ thống Edge AI trong IoT.....	29
III.3.1. Quy trình tổng quát.....	29
III.3.2. Bước 1 – Thu thập dữ liệu	29
III.3.3. Bước 2 – Tiền xử lý dữ liệu.....	30
III.3.4. Bước 3 – AI Inference	30
III.3.5. Bước 4 – Ra quyết định	30
III.3.6. Bước 5 – Đồng bộ dữ liệu lên Cloud.....	31
III.4. Các mô hình triển khai Edge AI trong IoT.....	31
III.4.1. Edge-only AI	31
III.4.2. Cloud-assisted Edge AI	31
III.4.3. Federated Edge AI.....	31
III.6. Ứng dụng Edge AI trong IoT	31
III.6.1. Giám sát môi trường thông minh.....	31
III.6.2. Camera AI thông minh	32
III.6.3. Công nghiệp thông minh	32
III.6.4. Y tế thông minh	32
III.7. Thách thức trong triển khai Edge AI cho IoT	32
III.7.1. Hạn chế tài nguyên	32
III.7.2. Tiêu thụ năng lượng.....	32
III.7.3. Bảo mật.....	33
III.7.4. Quản lý số lượng lớn thiết bị.....	33

III.8. Xu hướng phát triển AIoT	33
III.9. Ưu điểm của Edge AI trong IoT.....	33
III.9.1. Giảm độ trễ (Low Latency)	33
III.9.2. Xử lý thời gian thực (Real-time Processing)	34
III.9.3. Giảm tải băng thông mạng.....	35
III.9.4. Tăng tính riêng tư và bảo mật dữ liệu	35
III.9.5. Hoạt động ngay cả khi mất Internet.....	36
III.9.6. Tiết kiệm chi phí Cloud	36
III.9.7. Tăng khả năng mở rộng hệ thống.....	36
III.10. Thách thức của Edge AI trong IoT	36
III.10.1. Hạn chế tài nguyên phần cứng	36
III.10.2. Tiêu thụ năng lượng.....	37
III.10.3. Khó tối ưu mô hình AI	37
III.10.4. So sánh tổng hợp ưu điểm và thách thức.....	39
III.10.5. Các giải pháp khắc phục thách thức	39
III.10.6. Xu hướng nghiên cứu tương lai.....	40
CHƯƠNG IV: XU HƯỚNG VÀ HƯỚNG NGHIÊN CỨU TRONG EDGE AI.....	42
IV.1. Tổng quan xu hướng phát triển Edge AI.....	42
IV.2. Xu hướng phát triển Edge AI	42
IV.2.1. TinyML và AI siêu nhẹ	42
IV.2.2. Edge Generative AI	43
IV.2.3. Federated Learning.....	44
IV.2.4. AI Accelerator thế hệ mới	45
IV.2.5. 5G/6G và Edge AI.....	45
IV.2.6. Collaborative Edge Intelligence	46
IV.2.7. Green Edge AI.....	46
IV.3. Các hướng nghiên cứu mở trong Edge AI	46
IV.3.1. Tối ưu mô hình AI cho Edge.....	46
IV.3.2. Edge AI cho mạng cảm biến không dây	46
IV.3.3. Edge AI cho công nghiệp thông minh.....	47
IV.3.4. Edge AI cho mô thông minh	47
IV.3.5. Edge AI trong y tế thông minh.....	48
IV.3.6. Autonomous Edge Systems.....	48

IV.4. Các thách thức nghiên cứu trong tương lai	48
IV.4.1. Giới hạn tài nguyên	48
IV.4.2. Quản lý hệ thống phân tán.....	48
IV.4.3. Đảm bảo bảo mật AI	48
IV.4.4. Explainable AI (XAI)	49
IV.4.5. Quantum Computing Threat.....	49
IV.5. Định hướng nghiên cứu đề xuất	49
IV.6. Dự báo tương lai của Edge AI.....	49
CHƯƠNG V: MÔ HÌNH ỨNG DỤNG “HỆ THỐNG EDGE AI GIÁM SÁT CHẤT LƯỢNG KHÔNG KHÍ TẠI KHU CHẾ BIẾN THAN”	50
V.1. Giới thiệu bài toán	50
V.2. Mục tiêu của hệ thống.....	51
V.3. Kiến trúc tổng thể hệ thống	51
V.3.1. Kiến trúc ba lớp	51
V.4. Thành phần phần cứng hệ thống.....	52
V.4.1. Vi điều khiển ESP32.....	52
V.4.2. Cảm biến bụi PMS5003	53
V.4.3. Cảm biến khí MQ135	54
V.4.4. Raspberry Pi / Jetson Nano.....	54
V.5. Thiết kế phần mềm hệ thống	54
V.5.1. Kiến trúc phần mềm.....	54
V.5.2. Giao thức MQTT	55
V.5.3. Firebase Realtime Database.....	56
V.6. Mô hình Edge AI trong hệ thống.....	56
V.6.1. Dữ liệu đầu vào.....	56
V.6.2. Tiền xử lý dữ liệu.....	56
V.6.3. AI Model.....	56
V.6.4. AI Inference tại Edge.....	56
V.7. Cơ chế cảnh báo tự động	57
V.7.1. Điều kiện cảnh báo	57
V.7.2. Các hình thức cảnh báo.....	57
V.8. Ứng dụng Android giám sát hệ thống	58
V.9. Đánh giá hiệu quả hệ thống	59
V.9.1. Ưu điểm	59

V.9.2. Hạn chế	59
V.9.3. So sánh với Cloud AI.....	59
V.10. Khả năng mở rộng hệ thống	59
V.11. Hướng phát triển tương lai.....	60
KẾT LUẬN	61
TÀI LIỆU THAM KHẢO	62

MỞ ĐẦU

1. Đặt vấn đề

Trong những năm gần đây, sự phát triển mạnh mẽ của Internet vạn vật (IoT – Internet of Things) đã tạo ra một hệ sinh thái kết nối khổng lồ giữa các thiết bị thông minh, cảm biến và hệ thống điều khiển. Theo Cisco Systems, số lượng thiết bị IoT toàn cầu dự kiến đạt hàng chục tỷ thiết bị, tạo ra lượng dữ liệu khổng lồ theo thời gian thực. Các hệ thống IoT hiện đại được ứng dụng rộng rãi trong:

- Thành phố thông minh
- Công nghiệp 4.0
- Giao thông thông minh
- Y tế số
- Nông nghiệp thông minh
- Giám sát môi trường

Tuy nhiên, các hệ thống IoT truyền thống chủ yếu dựa trên mô hình điện toán đám mây (Cloud Computing), trong đó toàn bộ dữ liệu được gửi lên máy chủ Cloud để xử lý và suy luận trí tuệ nhân tạo (AI). Mô hình này tồn tại nhiều hạn chế:

- Độ trễ cao do truyền dữ liệu qua Internet
- Tiêu tốn băng thông mạng
- Phụ thuộc vào kết nối Internet liên tục
- Nguy cơ rò rỉ dữ liệu riêng tư
- Chi phí lưu trữ và xử lý lớn

Bảng 0.1 - Sự gia tăng dữ liệu IoT và áp lực đối với hạ tầng Cloud

Thành phần	Đặc điểm	Tác động
Thiết bị IoT tăng nhanh	Hàng tỷ thiết bị kết nối	Gia tăng dữ liệu liên tục
Dữ liệu thời gian thực	Video, cảm biến, âm thanh	Yêu cầu xử lý tức thời
Xử lý tập trung trên Cloud	Phụ thuộc Internet	Tăng độ trễ
AI truyền thống	Cần GPU mạnh trên Cloud	Chi phí vận hành cao

1.2. Xu hướng Edge AI trong hệ thống IoT

Để giải quyết các hạn chế của Cloud AI truyền thống, mô hình Edge AI đã ra đời và trở thành xu hướng nghiên cứu quan trọng trong lĩnh vực AIoT (Artificial Intelligence of Things).

Edge AI là mô hình đưa khả năng xử lý trí tuệ nhân tạo xuống gần nguồn sinh dữ liệu (edge devices), thay vì xử lý hoàn toàn trên Cloud. Các thiết bị Edge có thể thực hiện:

- Tiền xử lý dữ liệu
- Phân tích dữ liệu tại chỗ

- Nhận diện đối tượng
- Phát hiện bất thường
- Ra quyết định thời gian thực

Trong mô hình này:

- Dữ liệu được xử lý ngay tại thiết bị Edge
- Chỉ dữ liệu cần thiết mới gửi lên Cloud
- Hệ thống phản hồi nhanh hơn
- Tăng tính riêng tư và bảo mật

1.3. Vai trò của Edge AI trong IoT hiện đại

Sự kết hợp giữa Edge Computing và AI tạo ra nền tảng AIoT thông minh có khả năng:

- Xử lý dữ liệu thời gian thực
- Hoạt động độc lập khi mất Internet
- Tối ưu năng lượng và băng thông
- Nâng cao tính tự động hóa

Bảng 0.2 - So sánh giữa Cloud AI và Edge AI

Tiêu chí	Cloud AI	Edge AI
Vị trí xử lý	Trung tâm dữ liệu	Thiết bị biên
Độ trễ	Cao	Thấp
Băng thông	Tiêu tốn lớn	Tiết kiệm
Bảo mật dữ liệu	Dễ rò rỉ	An toàn hơn
Phản hồi thời gian thực	Hạn chế	Tốt
Phụ thuộc Internet	Cao	Thấp

1.4. Ứng dụng thực tiễn của Edge AI trong IoT

Hiện nay, Edge AI đang được triển khai mạnh mẽ trong nhiều lĩnh vực.

1.4.1. Thành phố thông minh

- Camera AI giao thông
- Phân tích mật độ phương tiện
- Nhận diện biển số xe
- Điều khiển tín hiệu giao thông thông minh

1.4.2. Công nghiệp thông minh

- Predictive Maintenance
- Giám sát dây chuyền sản xuất
- Robot tự hành
- Phát hiện lỗi thiết bị

1.4.3. Giám sát môi trường

- Phân tích chất lượng không khí
- Cảnh báo khí độc
- Giám sát bụi mịn PM2.5
- Cảnh báo cháy rừng

1.4.4. Y tế thông minh

- Thiết bị đeo thông minh
- Phân tích tín hiệu sinh học
- Theo dõi bệnh nhân từ xa

1.5. Tính cấp thiết của nghiên cứu

Trong bối cảnh chuyển đổi số và Công nghiệp 4.0, Edge AI đang trở thành công nghệ cốt lõi cho các hệ thống IoT thế hệ mới. Đặc biệt trong các lĩnh vực:

- Mô thông minh
- Nhà máy thông minh
- Giao thông thông minh
- Hệ thống giám sát môi trường thời gian thực

Các hệ thống này yêu cầu:

- Độ trễ cực thấp
- Tính ổn định cao
- Xử lý cục bộ
- Hoạt động liên tục ngay cả khi mất kết nối mạng

Do đó, nghiên cứu Edge AI không chỉ mang ý nghĩa học thuật mà còn có giá trị ứng dụng thực tiễn rất lớn.

1.6. Mục tiêu của báo cáo

Báo cáo tập trung vào các nội dung chính sau:

1. Trình bày tổng quan về Edge AI
2. Phân tích kiến trúc Edge AI trong IoT
3. Đánh giá ưu điểm và thách thức
4. Giới thiệu các nền tảng và công nghệ phổ biến
5. Phân tích các ứng dụng thực tiễn
6. Đề xuất hướng nghiên cứu trong tương lai

1.7. Phương pháp nghiên cứu

Các phương pháp được sử dụng gồm:

- Tổng hợp tài liệu khoa học
- Phân tích mô hình Edge AI

- So sánh các nền tảng AIoT
- Khảo sát ứng dụng thực tế
- Đánh giá kiến trúc triển khai

CHƯƠNG I: TỔNG QUAN VỀ EDGE AI

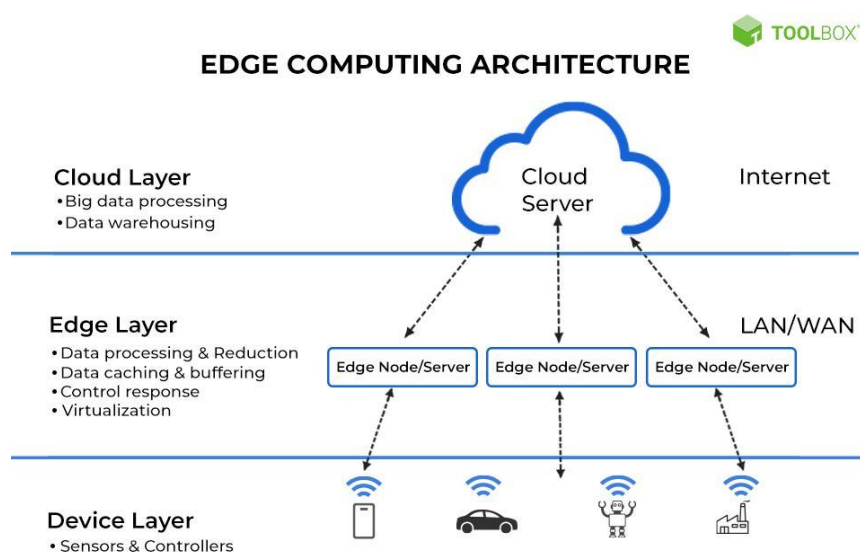
I.1. Khái niệm về Edge AI

I.1.1. Khái niệm Edge Computing

Edge Computing là mô hình điện toán phân tán trong đó việc xử lý dữ liệu được thực hiện gần nguồn sinh dữ liệu thay vì truyền toàn bộ dữ liệu đến trung tâm dữ liệu hoặc Cloud Server.

Theo IEEE, Edge Computing giúp:

- Giảm độ trễ truyền dữ liệu
- Tiết kiệm băng thông mạng
- Tăng khả năng xử lý thời gian thực
- Cải thiện tính riêng tư dữ liệu



Hình 1.1 – Kiến trúc Edge Computing

Trong hệ thống IoT truyền thống:

- Thiết bị cảm biến chỉ thu thập dữ liệu
- Dữ liệu được gửi toàn bộ lên Cloud
- Cloud thực hiện phân tích và phản hồi

Mô hình này phù hợp với các ứng dụng không yêu cầu phản hồi tức thời. Tuy nhiên đối với:

- Xe tự hành
- Robot công nghiệp
- Camera AI
- Giám sát khí độc
- Hệ thống cảnh báo cháy

thì độ trễ của Cloud Computing trở thành vấn đề nghiêm trọng.

1.1.2. Khái niệm Artificial Intelligence (AI)

Trí tuệ nhân tạo (AI) là lĩnh vực nghiên cứu cho phép máy tính mô phỏng khả năng tư duy và học tập của con người thông qua:

- Machine Learning
- Deep Learning
- Computer Vision
- Natural Language Processing

Các mô hình AI hiện đại có khả năng:

- Nhận diện hình ảnh
- Phân loại dữ liệu
- Dự đoán xu hướng
- Phát hiện bất thường
- Tự động ra quyết định

1.1.3. Khái niệm Edge AI

Edge AI là sự kết hợp giữa Edge Computing và Artificial Intelligence, trong đó các mô hình AI được triển khai trực tiếp trên các thiết bị Edge nhằm xử lý dữ liệu cục bộ mà không cần phụ thuộc hoàn toàn vào Cloud.

Định nghĩa:

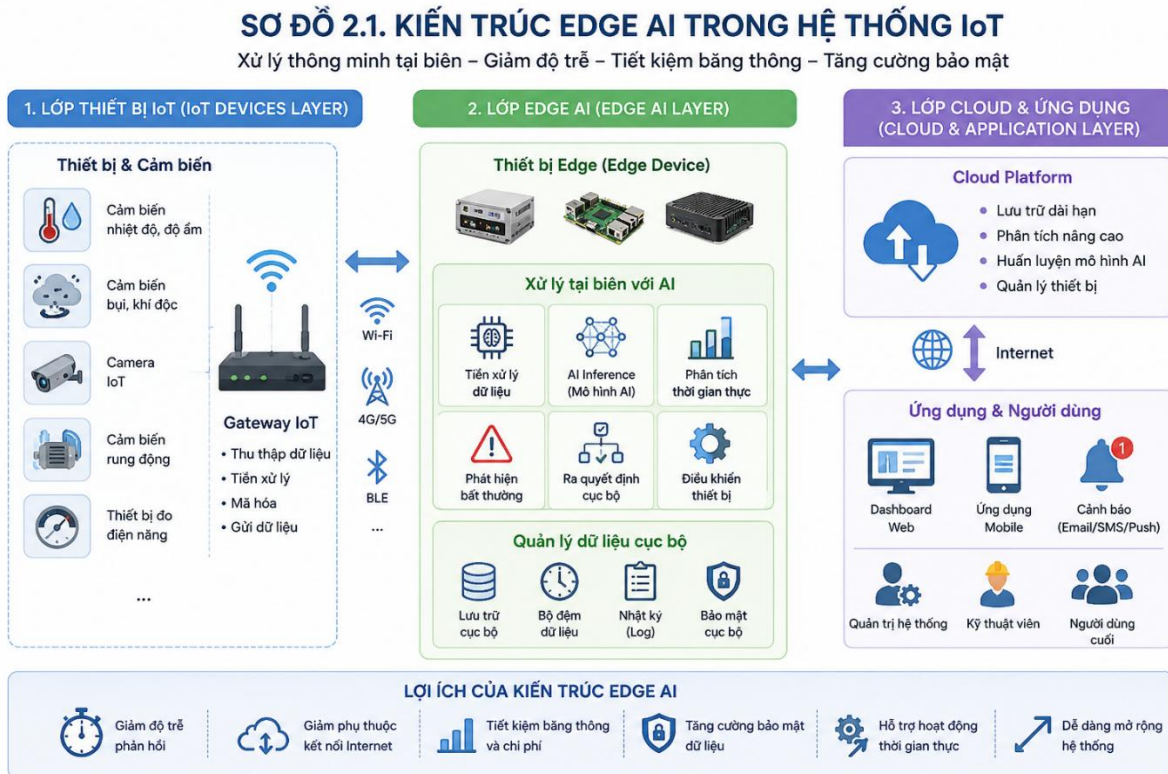
Edge AI là phương pháp triển khai và thực thi các thuật toán AI trực tiếp trên thiết bị biên nhằm cung cấp khả năng suy luận thời gian thực với độ trễ thấp.

Các thiết bị Edge AI có thể bao gồm:

- Smartphone
- Raspberry Pi
- NVIDIA Jetson
- Camera AI
- Gateway IoT
- Thiết bị nhúng AI

I.2. Kiến trúc tổng quát của Edge AI trong IoT

I.2.1. Mô hình kiến trúc tổng quát



Hình 1.2 - Kiến trúc Edge AI trong hệ thống IoT

I.2.2. Các thành phần chính

a) Device Layer

Bao gồm:

- Cảm biến nhiệt độ
- Cảm biến khí
- Camera
- Microphone
- Thiết bị IoT

Nhiệm vụ:

- Thu thập dữ liệu môi trường
- Gửi dữ liệu tới Edge Device

b) Edge Layer

Đây là thành phần quan trọng nhất của Edge AI. Chức năng:

- Tiền xử lý dữ liệu
- Chạy mô hình AI
- Phân tích thời gian thực
- Điều khiển cục bộ

Các nền tảng phổ biến:

- NVIDIA Jetson Nano

- Raspberry Pi Foundation Raspberry Pi
- Google Coral TPU
- ESP32 AI

c) Cloud Layer

Nhiệm vụ:

- Lưu trữ dữ liệu lớn
- Huấn luyện AI model
- Quản lý thiết bị IoT
- Đồng bộ hệ thống

I.3. Nguyên lý hoạt động của Edge AI

Quy trình xử lý dữ liệu

Bước 1: Thu thập dữ liệu

Cảm biến IoT thu thập:

- Hình ảnh
- Video
- Dữ liệu môi trường
- Âm thanh

Bước 2: Tiền xử lý tại Edge

- Loại bỏ nhiễu
- Chuẩn hóa dữ liệu
- Giảm kích thước dữ liệu

Bước 3: AI Inference

Thiết bị Edge chạy mô hình AI:

- CNN
- YOLO
- TinyML
- TensorFlow Lite

Bước 4: Ra quyết định

Ví dụ:

- Cảnh báo cháy
- Điều khiển quạt
- Phát hiện xâm nhập
- Cảnh báo khí độc

Bước 5: Gửi dữ liệu lên Cloud

- Chỉ dữ liệu quan trọng mới được đồng bộ lên Cloud.

I.4. So sánh Cloud AI và Edge AI

Bảng 1.1 - So sánh hai mô hình AI

Tiêu chí	Cloud AI	Edge AI
----------	----------	---------

Vị trí xử lý	Trung tâm dữ liệu	Thiết bị biên
Độ trễ	Cao	Thấp
Băng thông	Lớn	Nhỏ
Bảo mật dữ liệu	Trung bình	Cao
Xử lý thời gian thực	Hạn chế	Tốt
Phụ thuộc Internet	Cao	Thấp
Tiêu thụ năng lượng	Tập trung	Phân tán
Khả năng mở rộng	Cao	Trung bình

I.5. Các nền tảng Edge AI phổ biến

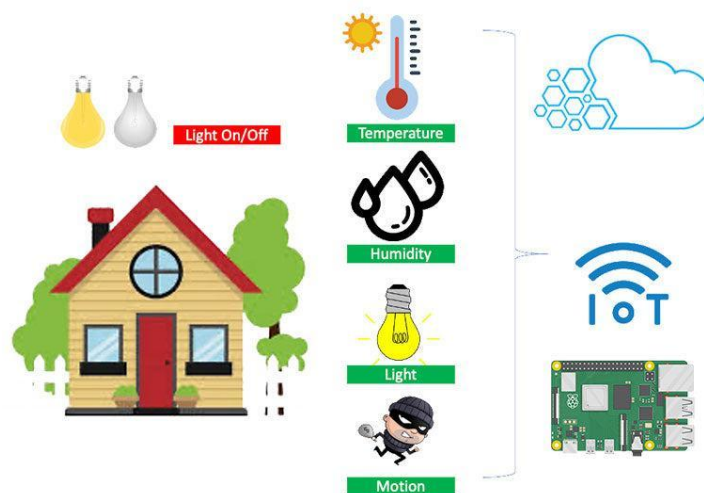
I.5.1. Raspberry Pi

Đặc điểm:

- Giá thành thấp
- Dễ phát triển
- Hỗ trợ Linux

Ứng dụng:

- Smart Home
- Camera AI
- AIoT nghiên cứu



I.5.2. NVIDIA Jetson

Đặc điểm:

- GPU mạnh
- Hỗ trợ CUDA
- Tăng tốc Deep Learning

Ứng dụng:

- Computer Vision
- Robot AI

- Xe tự hành



1.5.3. Google Coral TPU

Đặc điểm:

- TPU chuyên dụng cho AI
- Tốc độ inference cao
- Tiêu thụ điện năng thấp

Ứng dụng:

- TinyML
- AI Camera
- Smart Sensor

1.6. Các mô hình AI thường dùng tại Edge

1.6.1. CNN (Convolutional Neural Network)

Ứng dụng:

- Nhận diện ảnh
- Camera AI
- Phân loại vật thể

1.6.2. YOLO (You Only Look Once)

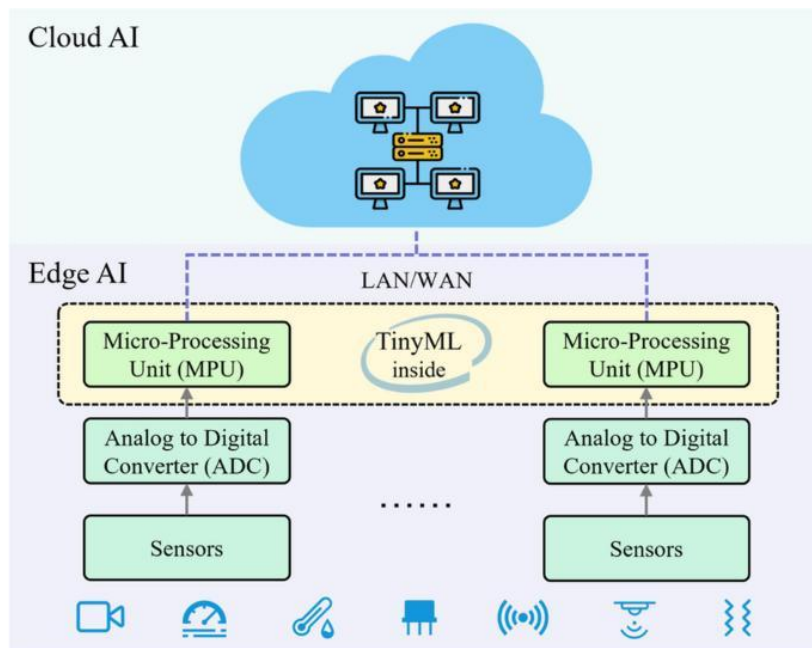
Ứng dụng:

- Object Detection
- Camera giao thông
- Drone AI

1.6.3. TinyML

TinyML là xu hướng triển khai AI trên các vi điều khiển công suất thấp như:

- ESP32
- Arduino Nano 33 BLE Sense



I.7. Ưu điểm của Edge AI

I.7.1. Độ trễ thấp

Do dữ liệu được xử lý cục bộ nên thời gian phản hồi nhanh hơn nhiều so với Cloud AI.

Ví dụ:

- Phát hiện cháy trong vài mili giây
- Xe tự hành phản ứng tức thời

I.7.2. Giảm tải băng thông

Chỉ dữ liệu quan trọng được gửi lên Cloud:

- Giảm lưu lượng mạng
- Giảm chi phí truyền tải

I.7.3. Tăng bảo mật dữ liệu

Dữ liệu nhạy cảm không cần gửi toàn bộ lên Internet.

Ví dụ:

- Camera an ninh
- Dữ liệu y tế
- Dữ liệu công nghiệp

I.8. Hạn chế của Edge AI

I.8.1. Hạn chế tài nguyên phần cứng

- CPU/GPU nhỏ
- RAM hạn chế
- Dung lượng lưu trữ thấp

1.8.2. Tiêu thụ năng lượng

Các mô hình Deep Learning yêu cầu:

- GPU mạnh
- Điện năng lớn

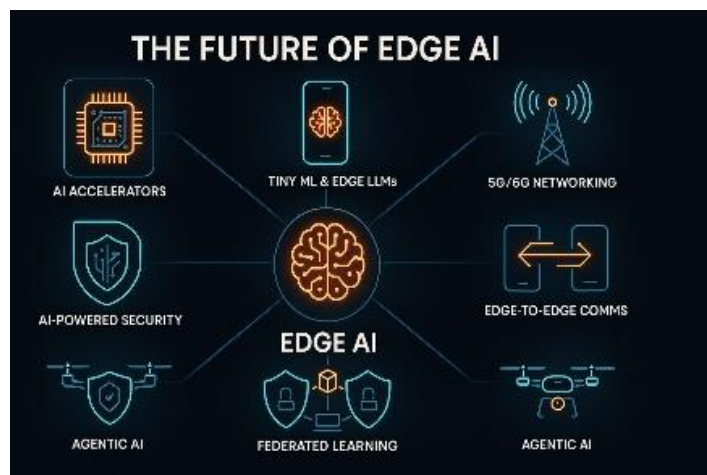
1.8.3. Khó cập nhật mô hình

Việc cập nhật hàng nghìn Edge Devices là thách thức lớn.

I.9. Xu hướng phát triển Edge AI

Các xu hướng nổi bật:

- TinyML
- Federated Learning
- Edge Generative AI
- AI Chip chuyên dụng
- 6G + Edge AI
- AIoT thông minh



CHƯƠNG II: CÁC CÔNG NGHỆ NỀN TẢNG TRONG EDGE AI

II.1. Tổng quan về công nghệ nền tảng Edge AI

Edge AI là sự kết hợp của nhiều công nghệ khác nhau bao gồm:

- Phần cứng nhúng hiệu năng cao
- Bộ tăng tốc AI (AI Accelerator)
- Framework học máy tối ưu
- Hệ điều hành nhúng
- Công nghệ truyền thông IoT
- Các kỹ thuật tối ưu mô hình AI

Những công nghệ này cho phép hệ thống:

- Xử lý dữ liệu thời gian thực
- Giảm độ trễ
- Hoạt động độc lập với Cloud
- Tối ưu điện năng tiêu thụ

II.2. Phần cứng trong Edge AI

II.2.1. Vai trò của phần cứng Edge AI

Phần cứng Edge AI là nền tảng thực thi các mô hình AI tại thiết bị biên. Khác với hệ thống Cloud sử dụng GPU công suất lớn, Edge AI yêu cầu:

- Kích thước nhỏ gọn
- Tiêu thụ điện năng thấp
- Xử lý thời gian thực
- Tích hợp AI accelerator

Các thành phần phần cứng chính gồm:

- CPU
- GPU
- TPU/NPU
- Bộ nhớ RAM
- Bộ nhớ lưu trữ
- Module truyền thông

II.2.2. Raspberry Pi

Raspberry Pi Foundation Raspberry Pi là nền tảng Edge phổ biến trong nghiên cứu AIoT.

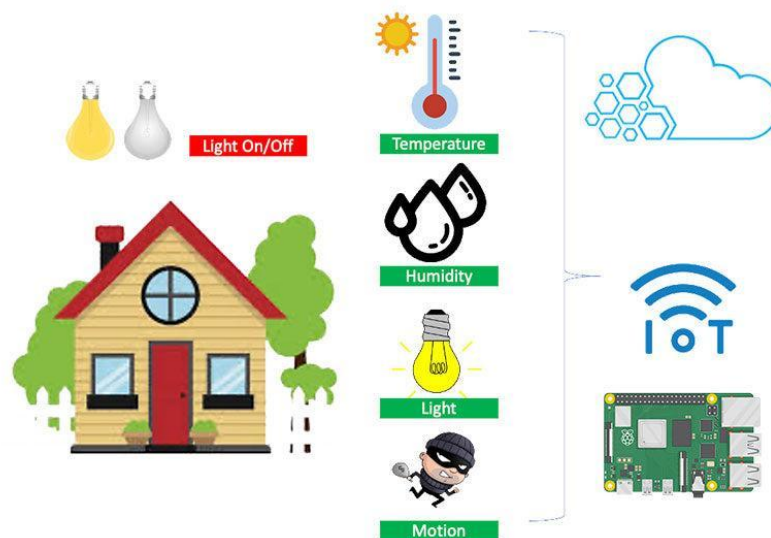
Đặc điểm:

- Giá thành thấp
- Hỗ trợ Linux
- Dễ lập trình

- Tích hợp GPIO cho IoT

Ứng dụng:

- Smart Home
- Camera AI
- Edge Gateway
- AIoT nghiên cứu



Bảng 2.1 - Thông số Raspberry Pi 5

Thành phần	Thông số
CPU	Quad-core ARM Cortex-A76
RAM	4GB / 8GB
GPU	VideoCore VII
Kết nối	WiFi, Bluetooth
Hệ điều hành	Raspberry Pi OS
Ứng dụng	IoT, AIoT, Gateway

II.2.3. NVIDIA Jetson

NVIDIA Jetson là nền tảng Edge AI mạnh mẽ dành cho Deep Learning và Computer Vision.

Đặc điểm:

- GPU CUDA mạnh
- Hỗ trợ TensorRT
- Tăng tốc AI inference
- Tối ưu cho Computer Vision

Các phiên bản phổ biến:

- Jetson Nano
- Jetson Xavier NX
- Jetson Orin



II.2.4. Google Coral TPU

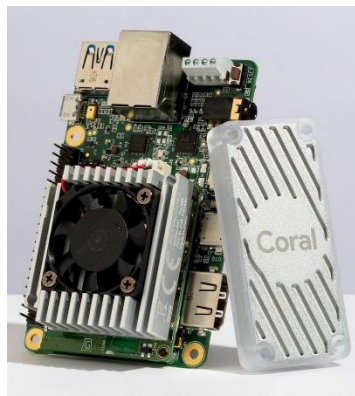
Google Coral TPU là bộ tăng tốc AI chuyên dụng cho Edge AI.

Đặc điểm:

- Tối ưu TensorFlow Lite
- Công suất thấp
- Inference tốc độ cao

Ứng dụng:

- Camera AI
- Smart Sensor
- TinyML



II.2.5. ESP32 AI và TinyML

ESP32 là vi điều khiển phổ biến trong AIoT nhờ:

- Giá rẻ
- Tích hợp WiFi/Bluetooth
- Hỗ trợ TinyML

Ứng dụng:

- Phân loại âm thanh
- Nhận diện cử chỉ
- Giám sát môi trường

II.3. Bộ tăng tốc AI (AI Accelerator)

II.3.1. Khái niệm

AI Accelerator là phần cứng chuyên dụng giúp tăng tốc:

- Deep Learning
- AI Inference
- Neural Network Processing

Các loại accelerator:

- GPU
- TPU
- NPU
- FPGA

II.3.2. GPU (Graphics Processing Unit)

GPU phù hợp với:

- Deep Learning
- Computer Vision
- Parallel Computing

Ưu điểm:

- Hiệu năng cao
- Hỗ trợ CUDA/OpenCL

Hạn chế:

- Tiêu thụ điện năng lớn

II.3.3. TPU (Tensor Processing Unit)

TPU được thiết kế riêng cho TensorFlow.

Đặc điểm:

- Tốc độ AI inference cao
- Tiêu thụ điện năng thấp
- Tối ưu ma trận tensor

II.3.4. NPU (Neural Processing Unit)

NPU là bộ xử lý AI chuyên dụng trong:

- Smartphone
- Camera AI
- IoT Edge Devices

Ví dụ:

- Apple Neural Engine
- Huawei Ascend
- Qualcomm AI Engine

II.4. Framework và phần mềm Edge AI

II.4.1. TensorFlow Lite

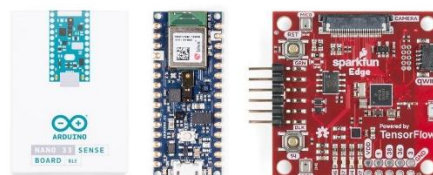
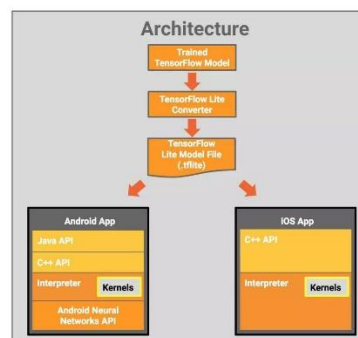
Google TensorFlow Lite là framework phổ biến cho Edge AI.

Đặc điểm:

- Tối ưu cho mobile/embedded
- Hỗ trợ quantization
- Kích thước nhỏ

Kiến trúc TensorFlow Lite

Tensorflow Lite



Quy trình triển khai:

1. Huấn luyện model
2. Chuyển đổi sang .tflite
3. Tối ưu model
4. Deploy tại Edge

II.4.2. PyTorch Mobile

Meta Platforms PyTorch Mobile hỗ trợ triển khai AI trên:

- Android
- iOS
- Embedded Linux

Ưu điểm:

- Dễ phát triển
- Hỗ trợ dynamic graph

II.4.3. OpenVINO

Intel OpenVINO là toolkit tối ưu AI cho CPU Intel.

Chức năng:

- Tăng tốc inference
- Tối ưu mô hình
- Hỗ trợ Computer Vision

II.4.4. Edge Impulse

Edge Impulse là nền tảng phát triển TinyML phổ biến.

Hỗ trợ:

- Thu thập dữ liệu
- Huấn luyện AI
- Triển khai TinyML
- Tự động sinh mã nhúng

II.5. Các mô hình AI sử dụng tại Edge

II.5.1. CNN (Convolutional Neural Network)

CNN chuyên xử lý:

- Hình ảnh
- Video
- Camera AI

Ứng dụng:

- Face Recognition
- Object Detection
- Medical Imaging

II.5.2. YOLO

YOLO là thuật toán Object Detection thời gian thực.

Đặc điểm:

- Tốc độ cao
- Độ chính xác tốt
- Phù hợp Edge AI

II.5.3. TinyML

TinyML cho phép triển khai AI trên:

- Vi điều khiển
- Thiết bị công suất thấp

Ưu điểm:

- Điện năng thấp
- Chi phí thấp
- Hoạt động offline

II.6. Công nghệ truyền thông trong Edge AI

II.6.1. MQTT

MQTT là giao thức phổ biến trong IoT:

- Lightweight
- Publish/Subscribe
- Độ trễ thấp

II.6.2. CoAP

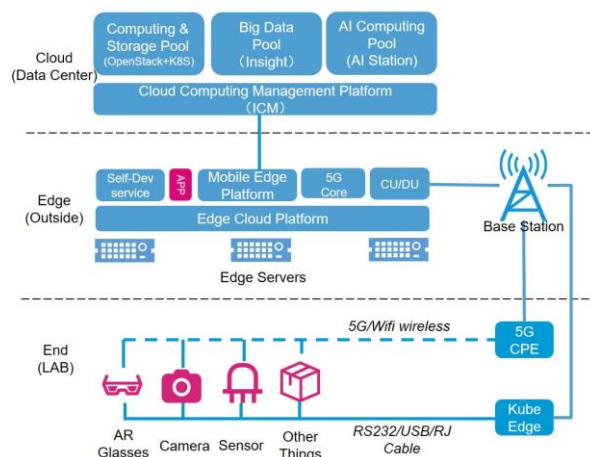
CoAP phù hợp với:

- Sensor Networks
- Embedded Devices

II.6.3. 5G và 6G

Mạng 5G hỗ trợ:

- Ultra-low latency
- Massive IoT
- Edge Computing



II.7. Các kỹ thuật tối ưu mô hình AI

II.7.1. Quantization

Mục tiêu:

- Giảm kích thước model
- Tăng tốc inference

Ví dụ:

- Float32 → INT8

II.7.2. Pruning

Loại bỏ:

- Neuron không cần thiết
- Trọng số dư thừa

II.7.3. Knowledge Distillation

Sử dụng:

- Teacher Model
- Student Model

để giảm kích thước mạng AI.

II.7.4. Model Compression

Nén mô hình nhằm:

- Giảm RAM
- Giảm bộ nhớ lưu trữ

II.8. Bảo mật trong nền tảng Edge AI

Các nguy cơ:

- Tấn công thiết bị Edge
- Đánh cắp model AI
- Giả mạo dữ liệu cảm biến

Giải pháp:

- Secure Boot
- TPM
- Federated Learning
- Blockchain

II.9. Xu hướng công nghệ Edge AI

Các xu hướng nổi bật:

- TinyML
- AI Chip chuyên dụng
- Edge Generative AI
- Federated Learning
- Collaborative Edge Intelligence

CHƯƠNG III: EDGE AI TRONG HỆ THỐNG IoT

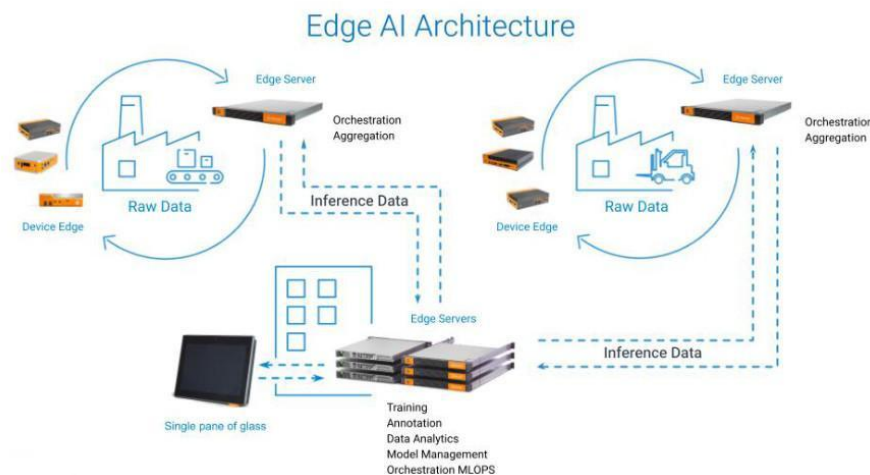
III.1. Tổng quan về tích hợp Edge AI và IoT

Sự kết hợp giữa Edge AI và Internet of Things (IoT) đã hình thành nên mô hình AIoT (Artificial Intelligence of Things), cho phép các thiết bị thông minh:

- Thu thập dữ liệu môi trường
- Phân tích dữ liệu cục bộ
- Ra quyết định thời gian thực
- Điều khiển tự động

Khác với hệ thống IoT truyền thống phụ thuộc hoàn toàn vào Cloud, Edge AI giúp đưa khả năng trí tuệ nhân tạo xuống gần nguồn dữ liệu hơn, từ đó:

- Giảm độ trễ
- Tăng tốc phản hồi
- Giảm tải mạng
- Nâng cao bảo mật dữ liệu



III.2. Kiến trúc tích hợp Edge AI trong IoT

III.2.1. Kiến trúc ba lớp (Three-layer Architecture)

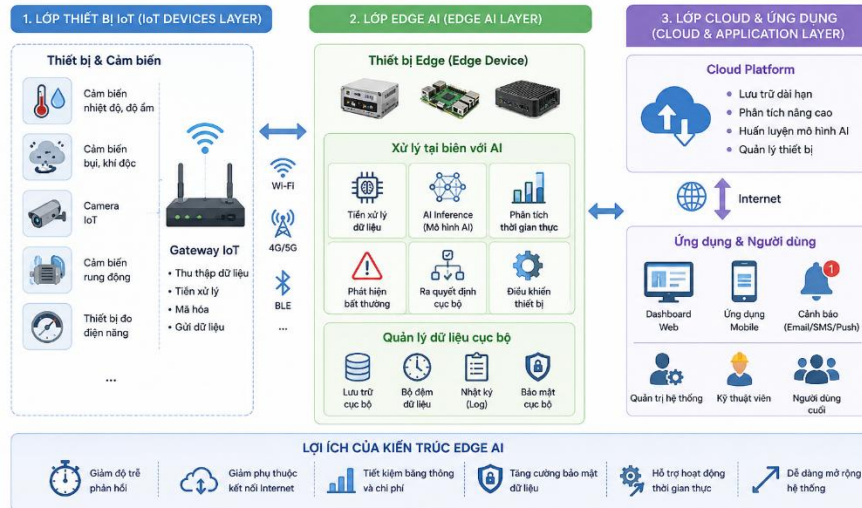
Kiến trúc phổ biến nhất của Edge AI trong IoT gồm ba lớp:

1. Device Layer
2. Edge Layer
3. Cloud Layer

Kiến trúc tổng quát AIoT gồm 3 lớp được thể hiện trên Hình 3.1.

SƠ ĐỒ 2.1. KIẾN TRÚC EDGE AI TRONG HỆ THỐNG IoT

Xử lý thông minh tại biên – Giảm độ trễ – Tiết kiệm băng thông – Tăng cường bảo mật



Hình 3.1 – Kiến trúc tổng quát AIIoT

III.2.2. Device Layer

Đây là lớp thấp nhất trong hệ thống AIIoT.

Thành phần:

- Cảm biến nhiệt độ
- Cảm biến khí
- Camera
- Microphone
- Thiết bị đo bụi PM2.5
- Thiết bị đeo thông minh

Chức năng:

- Thu thập dữ liệu môi trường
- Gửi dữ liệu đến Edge Gateway

Đặc điểm:

- Công suất thấp
- Tài nguyên hạn chế
- Hoạt động liên tục

III.2.3. Edge Layer

Edge Layer là trung tâm xử lý thông minh của hệ thống AIIoT.

Chức năng:

- Tiền xử lý dữ liệu
- Chạy mô hình AI
- Phân tích thời gian thực
- Điều khiển cục bộ
- Cảnh báo tức thời

Thiết bị phổ biến:

- NVIDIA Jetson Nano

- Raspberry Pi Foundation Raspberry Pi
- Google Coral TPU
- Industrial Edge Gateway

Ví dụ:

- Camera AI phát hiện cháy
- Nhận diện người xâm nhập
- Giám sát khí độc trong hầm lò

III.2.4. Cloud Layer

Cloud Layer đóng vai trò:

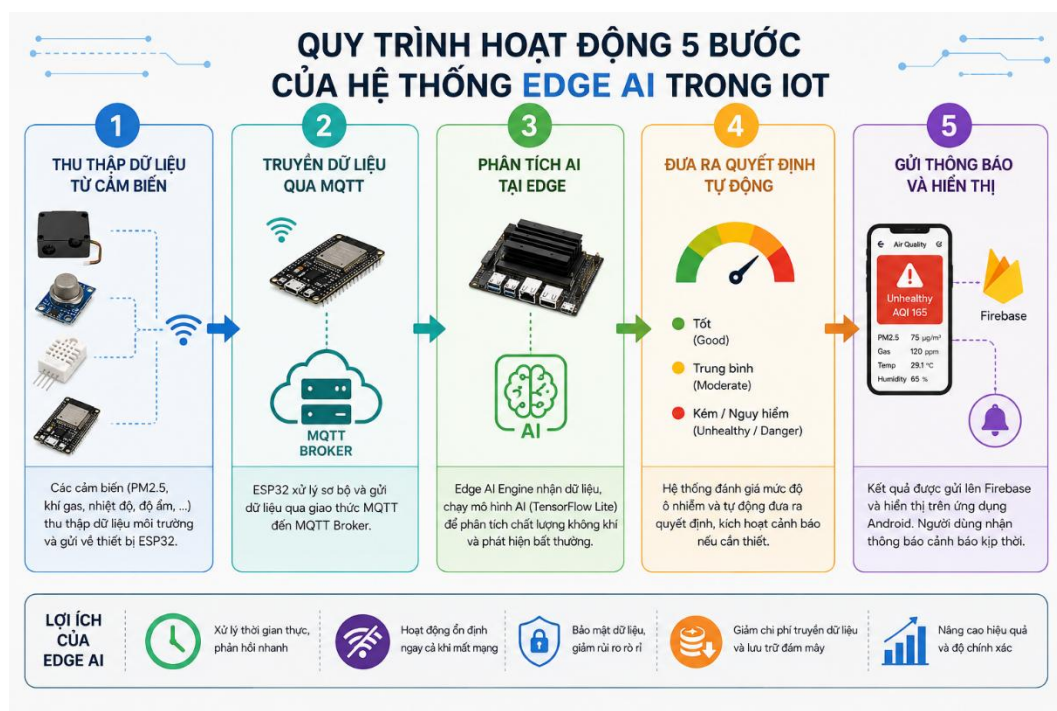
- Lưu trữ dữ liệu lớn
- Huấn luyện mô hình AI
- Đồng bộ hệ thống
- Quản lý thiết bị

Nền tảng phổ biến:

- Amazon Web Services AWS IoT
- Microsoft Azure IoT
- Google Cloud IoT
- Firebase

III.3. Quy trình hoạt động của hệ thống Edge AI trong IoT

III.3.1. Quy trình tổng quát



III.3.2. Bước 1 – Thu thập dữ liệu

Dữ liệu được thu thập từ:

- Camera
- Cảm biến khí
- Cảm biến nhiệt độ
- Cảm biến độ ẩm
- Microphone

Ví dụ:

Trong hệ thống giám sát môi trường:

- MQ135 đo khí độc
- PMS5003 đo bụi PM2.5
- DHT22 đo nhiệt độ và độ ẩm

III.3.3. Bước 2 – Tiền xử lý dữ liệu

Tại Edge Device, dữ liệu được:

- Chuẩn hóa
- Lọc nhiễu
- Giảm kích thước
- Trích xuất đặc trưng

Mục đích:

- Giảm tải xử lý
- Tăng độ chính xác AI
- Tiết kiệm tài nguyên

III.3.4. Bước 3 – AI Inference

Edge Device chạy mô hình AI:

- CNN
- YOLO
- TinyML
- TensorFlow Lite

Ví dụ:

Camera AI:

- Phát hiện người
- Nhận diện vật thể
- Phát hiện cháy

III.3.5. Bước 4 – Ra quyết định

Hệ thống Edge AI có thể:

- Kích hoạt còi cảnh báo
- Bật quạt thông gió
- Gửi SMS cảnh báo
- Điều khiển relay

Ví dụ:

Nếu nồng độ bụi PM2.5 vượt ngưỡng:

- Hệ thống tự động bật quạt hút bụi
- Gửi cảnh báo lên smartphone

III.3.6. Bước 5 – Đồng bộ dữ liệu lên Cloud

Chỉ dữ liệu quan trọng được gửi lên Cloud:

- Giảm băng thông
- Giảm chi phí lưu trữ
- Tăng bảo mật dữ liệu

III.4. Các mô hình triển khai Edge AI trong IoT

III.4.1. Edge-only AI

Toàn bộ AI chạy tại Edge Device.

Ưu điểm:

- Độ trễ thấp
- Hoạt động offline

Hạn chế:

- Hạn chế tài nguyên

III.4.2. Cloud-assisted Edge AI

Một phần xử lý thực hiện tại Edge, phần còn lại trên Cloud.

Ưu điểm:

- Cân bằng hiệu năng
- Hỗ trợ mô hình AI lớn

III.4.3. Federated Edge AI

Các Edge Devices cùng học AI mà không chia sẻ dữ liệu thô.

Ưu điểm:

- Tăng bảo mật
- Giảm rò rỉ dữ liệu

III.6. Ứng dụng Edge AI trong IoT

III.6.1. Giám sát môi trường thông minh

Ứng dụng:

- Phát hiện khí độc
- Giám sát bụi PM2.5
- Cảnh báo cháy

Ví dụ:

Hệ thống AIoT:

- ESP32 + MQ135
- Edge AI phân tích chất lượng không khí
- Android App hiển thị cảnh báo



III.6.2. Camera AI thông minh

Ứng dụng:

- Nhận diện khuôn mặt
- Phát hiện xâm nhập
- Giám sát giao thông

III.6.3. Công nghiệp thông minh

Ứng dụng:

- Predictive Maintenance
- Robot tự hành
- Giám sát máy móc

III.6.4. Y tế thông minh

Ứng dụng:

- Wearable Devices
- ECG Monitoring
- Health Analytics

III.7. Thách thức trong triển khai Edge AI cho IoT

III.7.1. Hạn chế tài nguyên

- CPU yếu
- RAM nhỏ
- Dung lượng thấp

III.7.2. Tiêu thụ năng lượng

Deep Learning yêu cầu:

- GPU mạnh

- Điện năng cao

III.7.3. Bảo mật

Nguy cơ:

- Tấn công Edge Device
- Đánh cắp dữ liệu
- Giả mạo thiết bị

III.7.4. Quản lý số lượng lớn thiết bị

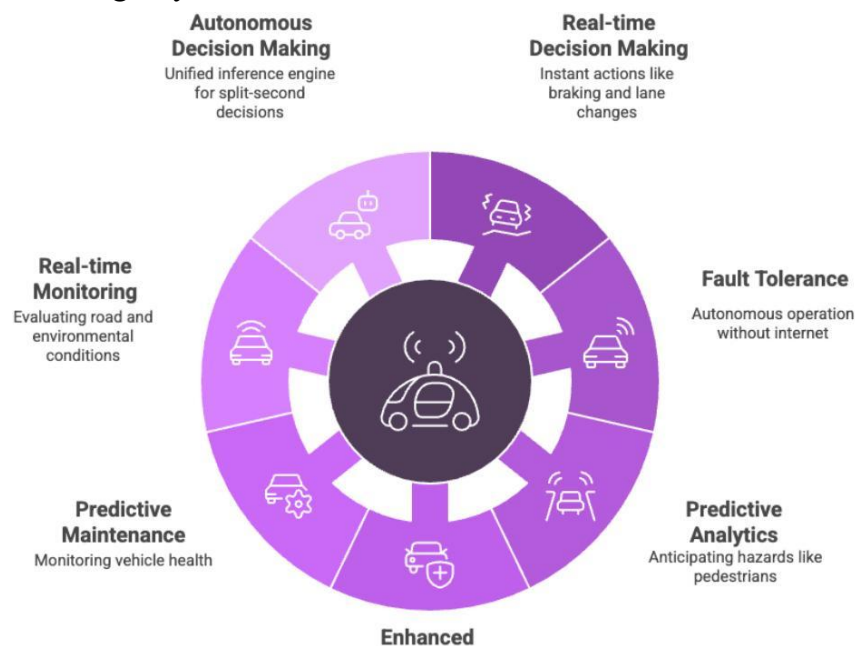
Hệ thống AIoT có thể gồm:

- Hàng nghìn sensor
- Nhiều Edge Nodes
- Phân tán địa lý rộng

III.8. Xu hướng phát triển AIoT

Các xu hướng nổi bật:

- TinyML
- Federated Learning
- Edge Generative AI
- 6G AIoT
- Autonomous Edge Systems



III.9. Ưu điểm của Edge AI trong IoT

III.9.1. Giảm độ trễ (Low Latency)

Một trong những ưu điểm quan trọng nhất của Edge AI là khả năng xử lý dữ liệu trực tiếp tại thiết bị biên thay vì gửi toàn bộ dữ liệu lên Cloud.

Trong mô hình Cloud AI:

Sensor → Internet → Cloud → AI Processing → Response

Trong mô hình Edge AI:

Sensor → Edge AI → Response

Điều này giúp:

- Giảm thời gian truyền dữ liệu
- Phản hồi gần như tức thời
- Tăng hiệu quả điều khiển

Ví dụ thực tế: Xe tự hành

Xe tự hành cần:

- Nhận diện vật cản
- Phanh khẩn cấp
- Điều hướng tức thời

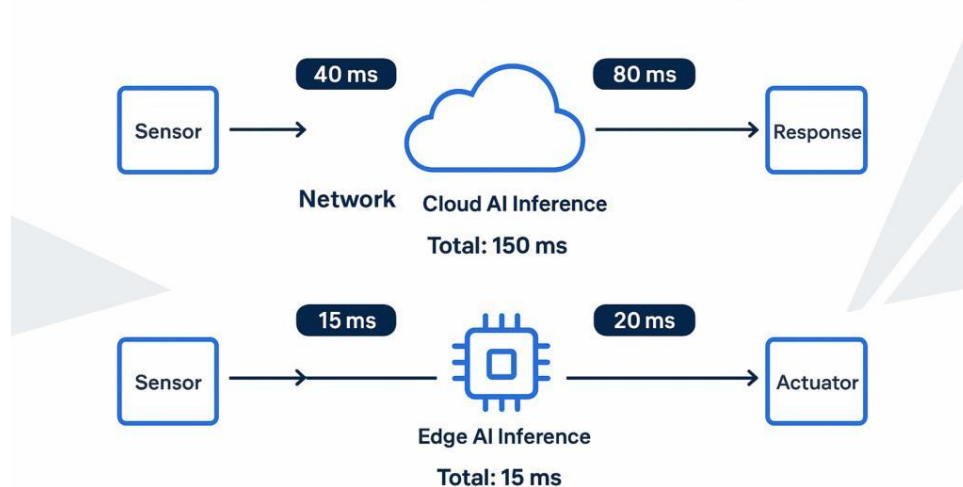
Nếu xử lý trên Cloud:

- Độ trễ vài trăm mili giây có thể gây tai nạn.

Với Edge AI:

- AI xử lý trực tiếp trên xe
- Phản hồi trong vài mili giây

Cloud vs. Edge AI Latency



III.9.2. Xử lý thời gian thực (Real-time Processing)

Edge AI cho phép:

- Giám sát liên tục
- Phân tích dữ liệu streaming
- Cảnh báo ngay lập tức

Ứng dụng:

- Camera AI

- Robot công nghiệp
- Giám sát khí độc
- Hệ thống chống cháy

Ví dụ:

Trong hệ thống giám sát hầm lò:

- Sensor MQ135 phát hiện khí độc
- Edge AI phân tích dữ liệu
- Relay kích hoạt quạt thông gió ngay lập tức

III.9.3. Giảm tải băng thông mạng

Trong hệ thống IoT truyền thống:

- Toàn bộ dữ liệu gửi lên Cloud
- Tốn băng thông lớn

Edge AI chỉ gửi:

- Dữ liệu quan trọng
- Kết quả phân tích
- Metadata

Ví dụ:

Camera giám sát:

- Không gửi toàn bộ video
- Chỉ gửi khi phát hiện bất thường

Bảng 3.1 - So sánh lưu lượng dữ liệu

Hệ thống	Dữ liệu truyền
Cloud AI	Video liên tục
Edge AI	Chỉ sự kiện bất thường

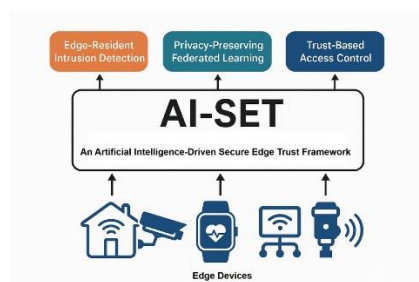
III.9.4. Tăng tính riêng tư và bảo mật dữ liệu

Dữ liệu nhạy cảm được xử lý cục bộ:

- Không cần truyền toàn bộ lên Internet
- Giảm nguy cơ rò rỉ dữ liệu

Ví dụ:

- Camera bệnh viện
- Thiết bị y tế
- Nhà máy thông minh



III.9.5. Hoạt động ngay cả khi mất Internet

Edge AI vẫn có thể:

- Chạy AI inference
- Điều khiển cục bộ
- Ra quyết định offline

Rất quan trọng trong:

- Hàm lò khai thác
- Khu vực vùng sâu
- Nhà máy tự động

III.9.6. Tiết kiệm chi phí Cloud

Edge AI giúp giảm:

- Chi phí truyền dữ liệu
- Chi phí lưu trữ Cloud
- Chi phí xử lý GPU trên Data Center

III.9.7. Tăng khả năng mở rộng hệ thống

Hệ thống phân tán Edge:

- Giảm tải trung tâm
- Dễ mở rộng theo khu vực
- Tăng tính linh hoạt

III.10. Thách thức của Edge AI trong IoT

III.10.1. Hạn chế tài nguyên phần cứng

Các Edge Devices thường có:

- CPU yếu
- RAM nhỏ
- Bộ nhớ hạn chế
- Công suất thấp

Trong khi đó mô hình AI hiện đại:

- Deep Learning
- Computer Vision
- LLMs

đòi hỏi:

- GPU mạnh
- Bộ nhớ lớn

Ví dụ:

NVIDIA Jetson Nano chỉ có:

- RAM 4GB
- GPU hạn chế

nên khó triển khai:

- Large Language Models
- Video Analytics độ phân giải cao

III.10.2. Tiêu thụ năng lượng

AI inference tiêu tốn nhiều điện năng.

Đặc biệt:

- CNN
- YOLO
- Computer Vision

làm tăng:

- Nhiệt độ hệ thống
- Tiêu hao pin
- Chi phí vận hành

III.10.3. Khó tối ưu mô hình AI

Mô hình AI phải được:

- Giảm kích thước
- Tối ưu inference
- Giảm số tham số

Các kỹ thuật thường dùng:

- Quantization
- Pruning
- Knowledge Distillation

Ví dụ:

YOLOv8 bản đầy đủ:

- Khó chạy trên ESP32

Cần:

- TinyML
- TensorFlow Lite
- Phiên bản lightweight

III.10.3.1. Bảo mật thiết bị Edge

Thiết bị Edge dễ bị:

- Tấn công vật lý
- Đánh cắp firmware
- Chiếm quyền điều khiển

Các kiểu tấn công phổ biến

a) Malware Injection

- Chèn mã độc vào Edge Device.

b) Model Stealing

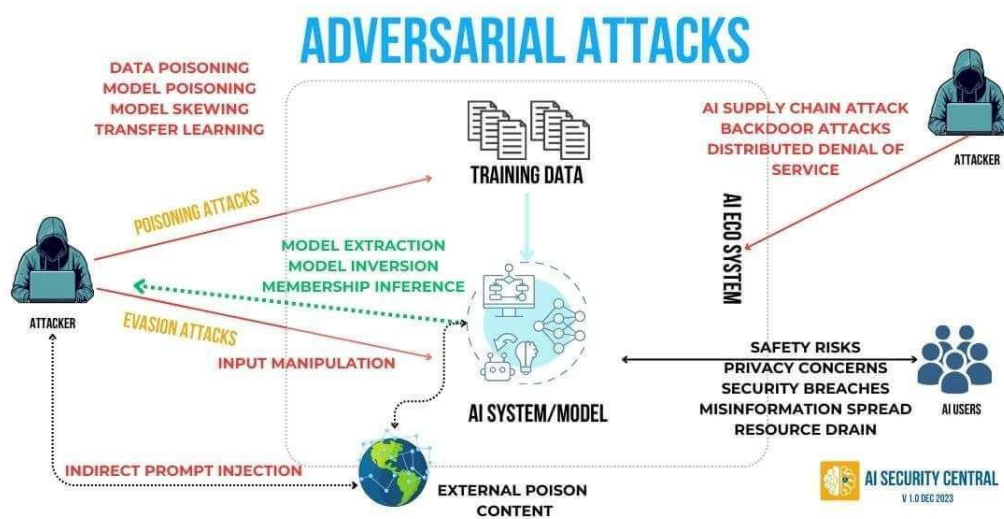
- Đánh cắp mô hình AI.

c) Data Poisoning

- Làm sai lệch dữ liệu huấn luyện.

d) Adversarial Attack

- Tấn công đánh lừa mô hình AI.



III.10.3.2. Quản lý số lượng lớn thiết bị IoT

Hệ thống AIoT có thể gồm:

- Hàng nghìn sensor
- Nhiều Edge Nodes
- Phân tán địa lý rộng

Khó khăn:

- Đồng bộ firmware
- Cập nhật AI model
- Quản lý lỗi hệ thống

III.10.3.3. Khó cập nhật mô hình AI

Mô hình AI cần:

- Retraining
- Fine-tuning
- Update liên tục

Trong hệ thống lớn:

- Quá trình cập nhật rất phức tạp

III.10.3.4. Khả năng tương thích hệ thống

Edge AI phải tích hợp với:

- Nhiều loại sensor
- Nhiều giao thức IoT
- Nhiều nền tảng Cloud

Điều này gây khó khăn:

- Chuẩn hóa hệ thống
- Tương thích phần mềm
- Đồng bộ dữ liệu

III.10.4. So sánh tổng hợp ưu điểm và thách thức

Bảng 3.2 - Tổng hợp đánh giá Edge AI

Nội dung	Ưu điểm	Thách thức
Độ trễ	Thấp	Phụ thuộc phần cứng
Băng thông	Tiết kiệm	Quản lý phân tán
Bảo mật	Tăng riêng tư	Dễ bị tấn công vật lý
AI Processing	Real-time	Giới hạn tài nguyên
Chi phí	Giảm Cloud	Tăng chi phí Edge
Năng lượng	Hiệu quả hơn Cloud	Pin hạn chế

III.10.5. Các giải pháp khắc phục thách thức

III.10.5.1. TinyML

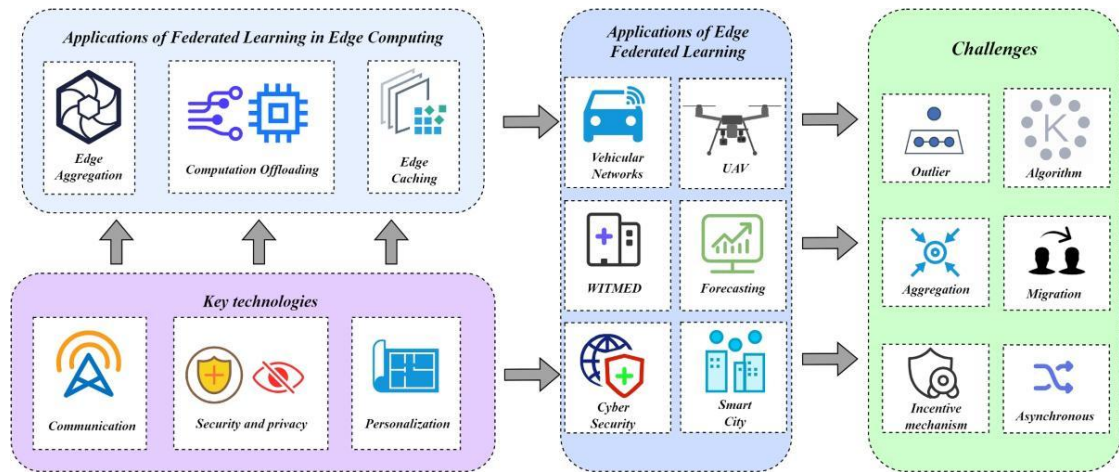
TinyML giúp:

- Triển khai AI trên vi điều khiển
- Giảm điện năng
- Giảm chi phí phần cứng

III.10.5.2. Federated Learning

Federated Learning cho phép:

- Huấn luyện phân tán
- Không chia sẻ dữ liệu thô
- Tăng bảo mật



III.10.5.3. AI Accelerator

Các AI Chip chuyên dụng:

- TPU
- NPU
- Edge GPU

giúp:

- Tăng tốc inference
- Giảm điện năng

III.10.5.4. Edge-Cloud Collaboration

Kết hợp:

- Edge xử lý real-time
- Cloud xử lý dữ liệu lớn

để tối ưu hiệu năng toàn hệ thống.

III.10.6. Xu hướng nghiên cứu tương lai

Các hướng nghiên cứu nổi bật:

- Green Edge AI
- Tiny Generative AI
- Edge LLM
- 6G AIoT
- Autonomous Edge Systems



CHƯƠNG IV: XU HƯỚNG VÀ HƯỚNG NGHIÊN CỨU TRONG EDGE AI

IV.1. Tổng quan xu hướng phát triển Edge AI

Trong bối cảnh:

- Dữ liệu IoT tăng trưởng bùng nổ
- Nhu cầu xử lý thời gian thực ngày càng cao
- AI được tích hợp vào mọi lĩnh vực

Edge AI đang chuyển từ mô hình hỗ trợ sang nền tảng trung tâm của hệ sinh thái AIoT.

Các xu hướng phát triển mới tập trung vào:

- AI hiệu năng cao công suất thấp
- AI phân tán
- Edge Intelligence
- TinyML
- Edge Generative AI
- 6G AIoT

IV.2. Xu hướng phát triển Edge AI

IV.2.1. TinyML và AI siêu nhẹ

TinyML là xu hướng triển khai mô hình học máy trên:

- Vi điều khiển
- Thiết bị công suất thấp
- Sensor thông minh

Các nền tảng phổ biến:

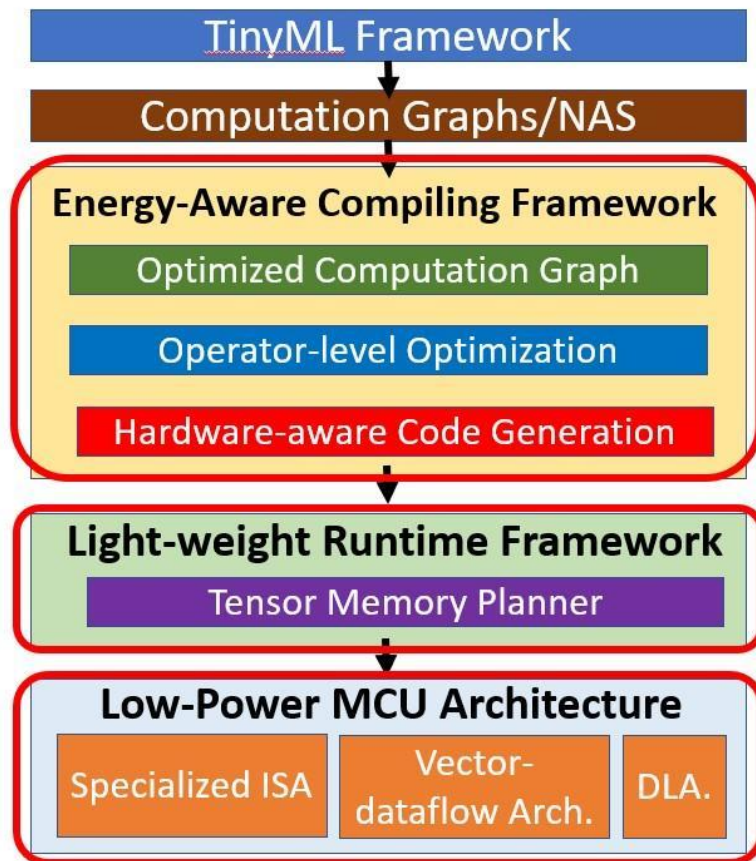
- ESP32
- Arduino Nano 33 BLE Sense
- STM32
- Coral Micro

Đặc điểm của TinyML

- Điện năng cực thấp
- Hoạt động offline
- Chi phí thấp
- Thời gian phản hồi nhanh

Ứng dụng:

- Nhận diện giọng nói
- Phát hiện rung động
- Giám sát khí độc
- Phân loại âm thanh



Xu hướng nghiên cứu TinyML

- Nén mô hình AI
- Quantization tối ưu
- Tiny Transformer
- Tiny Vision Models

IV.2.2. Edge Generative AI

Generative AI đang được đưa xuống Edge Devices nhằm:

- Giảm phụ thuộc Cloud
- Tăng tính riêng tư
- Giảm độ trễ

Các ứng dụng:

- Sinh văn bản cục bộ
- Voice Assistant offline
- AI chatbot tại Edge
- Sinh ảnh thời gian thực

Thách thức:

- Mô hình lớn

- Yêu cầu GPU mạnh
- Tiêu thụ điện năng cao

Xu hướng:

- Tiny LLM
- Quantized LLM
- Distilled Language Models



IV.2.3. Federated Learning

Federated Learning là mô hình huấn luyện AI phân tán:

- Không gửi dữ liệu thô lên Cloud
- Chỉ chia sẻ trọng số mô hình

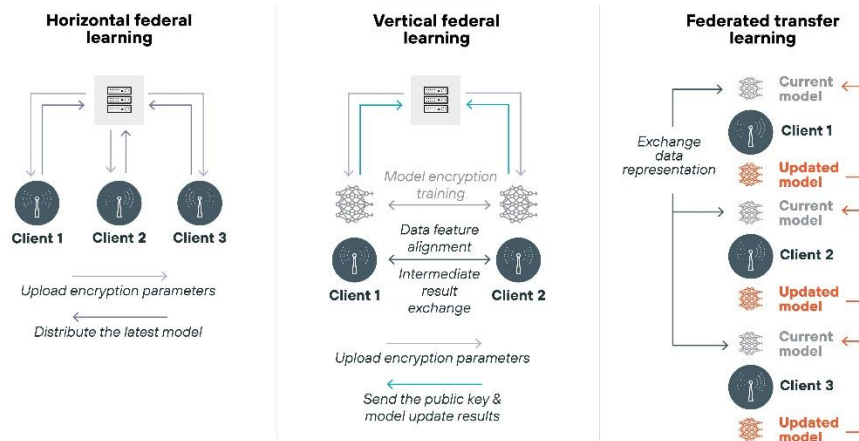
Ưu điểm:

- Bảo vệ quyền riêng tư
- Giảm lưu lượng mạng
- Phù hợp hệ thống AIoT lớn

Ứng dụng:

- Smart Healthcare
- Smart Factory
- Autonomous Vehicles

Categories of federated learning



IV.2.4. AI Accelerator thế hệ mới

Các AI Accelerator mới được thiết kế chuyên biệt cho Edge AI:

- TPU
- NPU
- Edge GPU
- FPGA AI

Xu hướng:

- Hiệu năng cao hơn
- Điện năng thấp hơn
- Tích hợp AI engine chuyên dụng

Ví dụ:

- NVIDIA Jetson Orin
- Google Edge TPU
- Intel Movidius
- Apple Neural Engine

IV.2.5. 5G/6G và Edge AI

Mạng 5G và tương lai là 6G đóng vai trò quan trọng trong AIoT.

Đặc điểm:

- Ultra-low latency
- Massive IoT
- Edge-native networking
- Intelligent communication

Ứng dụng:

- Xe tự hành
- Smart City
- Drone AI
- Robot công nghiệp

IV.2.6. Collaborative Edge Intelligence

Nhiều Edge Nodes cùng:

- Chia sẻ dữ liệu
- Chia sẻ tài nguyên
- Phối hợp AI inference

Mục tiêu:

- Tăng hiệu năng
- Giảm tải từng node
- Mở rộng hệ thống

IV.2.7. Green Edge AI

Green AI tập trung:

- Giảm điện năng
- Giảm phát thải carbon
- Tăng hiệu quả năng lượng

Hướng nghiên cứu:

- Low-power AI Chips
- Energy-aware AI
- Carbon-efficient Computing

IV.3. Các hướng nghiên cứu mở trong Edge AI

IV.3.1. Tối ưu mô hình AI cho Edge

Bài toán:

Mô hình AI hiện đại:

- Quá lớn
- Tiêu thụ tài nguyên cao

Hướng nghiên cứu:

- Model Compression
- Quantization
- Pruning
- Knowledge Distillation

Ví dụ:

YOLO Nano cho Edge Camera.

IV.3.2. Edge AI cho mạng cảm biến không dây

Wireless Sensor Networks (WSN):

- Công suất thấp
- Năng lượng hạn chế

Các vấn đề nghiên cứu:

- AI routing optimization
- Fault detection

- Distributed intelligence

IV.3.3. Edge AI cho công nghiệp thông minh

Hướng nghiên cứu:

- Predictive Maintenance
- Industrial Robotics
- Digital Twin
- Smart Manufacturing

Các bài toán phổ biến:

- Phát hiện lỗi máy móc
- Giám sát rung động
- AI Vision trong sản xuất

IV.3.4. Edge AI cho mỏ thông minh

Đây là hướng nghiên cứu rất tiềm năng trong:

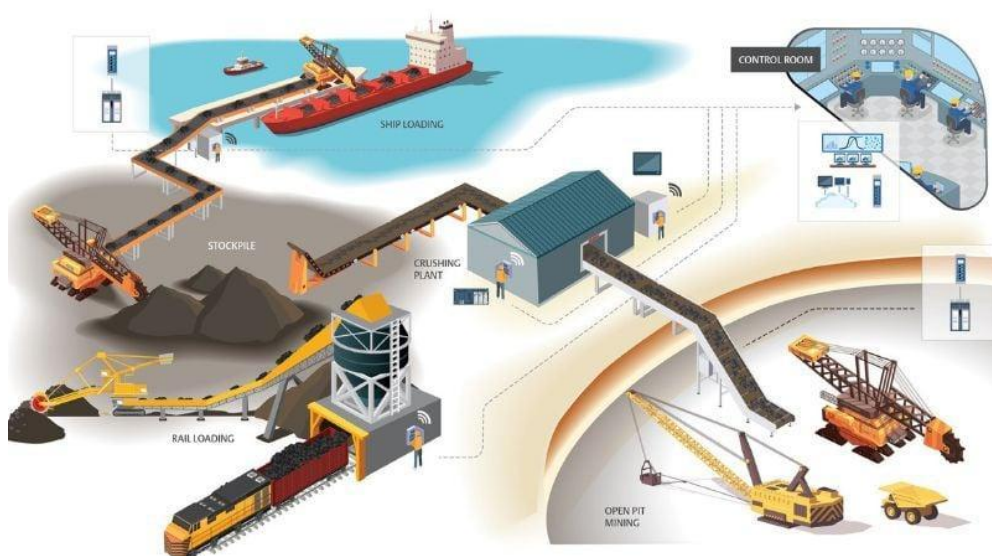
- Khai thác than
- Hàm lò
- Mỏ lộ thiên

Ứng dụng:

- Giám sát khí độc
- Cảnh báo bụi mịn
- Camera AI an toàn lao động
- Robot tự hành trong hầm lò

Các thách thức:

- Môi trường khắc nghiệt
- Mất kết nối mạng
- Tiêu thụ năng lượng



IV.3.5. Edge AI trong y tế thông minh

Hướng nghiên cứu:

- Wearable AI
- Remote Health Monitoring
- ECG Analysis
- AI Medical Imaging

Thách thức:

- Quyền riêng tư dữ liệu
- Độ chính xác AI
- Bảo mật thiết bị y tế

IV.3.6. Autonomous Edge Systems

Các hệ thống Edge AI tương lai sẽ:

- Tự học
- Tự thích nghi
- Tự ra quyết định

Ví dụ:

- Drone tự hành
- Robot AI
- Smart Vehicles

IV.4. Các thách thức nghiên cứu trong tương lai

IV.4.1. Giới hạn tài nguyên

Edge Devices vẫn bị hạn chế:

- CPU
- GPU
- RAM
- Pin

IV.4.2. Quản lý hệ thống phân tán

Hệ thống AIoT tương lai có thể gồm:

- Hàng triệu thiết bị
- Nhiều Edge Nodes
- Địa lý phân tán

IV.4.3. Đảm bảo bảo mật AI

Các nguy cơ:

- Adversarial Attack
- AI Malware
- Model Stealing

IV.4.4. Explainable AI (XAI)

AI cần:

- Giải thích được quyết định
- Tăng tính minh bạch
- Đảm bảo độ tin cậy

IV.4.5. Quantum Computing Threat

Máy tính lượng tử trong tương lai có thể:

- Phá vỡ thuật toán mã hóa hiện tại
- Gây rủi ro bảo mật lớn

IV.5. Định hướng nghiên cứu đề xuất

Các hướng nghiên cứu phù hợp tại Việt Nam

a) Edge AI giám sát môi trường

- Bụi PM2.5
- Khí độc
- Chất lượng không khí

b) AIoT cho mở thông minh

- Giám sát an toàn hầm lò
- Cảnh báo cháy nổ
- Robot kiểm tra tự động

c) TinyML cho nông nghiệp

- Giám sát độ ẩm
- Nhận diện sâu bệnh
- Tưới tiêu thông minh

d) Edge AI cho giao thông thông minh

- Camera AI giao thông
- Nhận diện biển số
- Phân tích mật độ xe

IV.6. Dự báo tương lai của Edge AI

Theo nhiều báo cáo của:

- International Data Corporation
- Gartner
- IEEE

Edge AI sẽ:

- Trở thành nền tảng cốt lõi của AIoT
- Được triển khai rộng trong công nghiệp
- Tích hợp mạnh với 6G
- Hỗ trợ hệ thống tự động thông minh

CHƯƠNG V: MÔ HÌNH ỨNG DỤNG “HỆ THỐNG EDGE AI GIÁM SÁT CHẤT LƯỢNG KHÔNG KHÍ TẠI KHU CHẾ BIẾN THAN”

V.1. Giới thiệu bài toán

Trong các khu chế biến than, quá trình:

- Nghiền than
- Sàng tuyển
- Vận chuyển
- Bốc dỡ vật liệu

thường phát sinh:

- Bụi mịn PM2.5/PM10
- Khí CO
- Khí CH₄
- SO₂
- NO₂

Các tác nhân này ảnh hưởng nghiêm trọng đến:

- Sức khỏe người lao động
- Môi trường sản xuất
- An toàn cháy nổ
- Thiết bị công nghiệp

Trong nhiều hệ thống truyền thống:

- Dữ liệu cảm biến được gửi toàn bộ lên Cloud
- Độ trễ xử lý cao
- Phụ thuộc Internet
- Khó phản ứng thời gian thực

Do đó, việc ứng dụng Edge AI trong giám sát chất lượng không khí là hướng tiếp cận hiệu quả nhằm:

- Phân tích dữ liệu tại chỗ
- Phát hiện bất thường tức thời
- Tự động cảnh báo
- Điều khiển hệ thống xử lý môi trường



V.2. Mục tiêu của hệ thống

Hệ thống được xây dựng nhằm:

- Giám sát liên tục chất lượng không khí
- Phát hiện bụi và khí độc vượt ngưỡng
- Phân tích dữ liệu bằng Edge AI
- Cảnh báo thời gian thực
- Điều khiển tự động hệ thống xử lý bụi

Các chức năng chính

a) Thu thập dữ liệu môi trường

- PM2.5
- PM10
- CO
- CH₄
- Nhiệt độ
- Độ ẩm

b) Phân tích Edge AI

- Phát hiện bất thường
- Dự đoán xu hướng ô nhiễm
- Phân loại mức độ nguy hiểm

c) Điều khiển tự động

- Bật quạt hút bụi
- Kích hoạt phun sương
- Cảnh báo còi/đèn

d) Giám sát từ xa

- Android App
- Dashboard Web
- Firebase Cloud

V.3. Kiến trúc tổng thể hệ thống

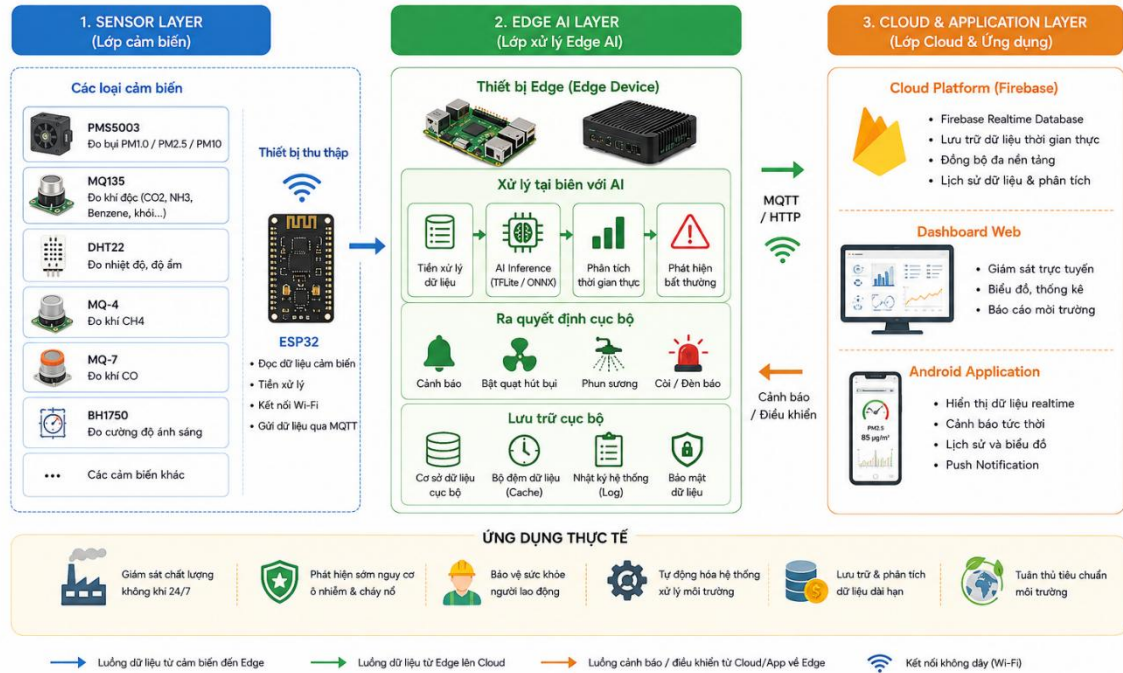
V.3.1. Kiến trúc ba lớp

Hệ thống gồm:

1. Sensor Layer
2. Edge AI Layer
3. Cloud & Application Layer

SƠ ĐỒ 9.1. KIẾN TRÚC HỆ THỐNG

Hệ thống Edge AI giám sát chất lượng không khí tại khu chế biến than



Hình 5.1 – Sơ đồ kiến trúc hệ thống

V.4. Thành phần phần cứng hệ thống

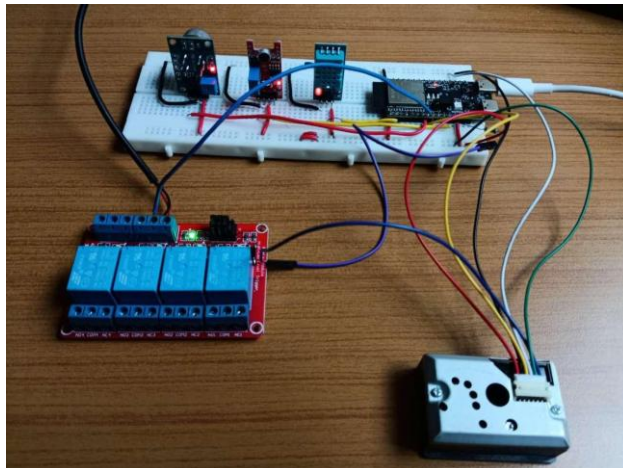
V.4.1. Vi điều khiển ESP32

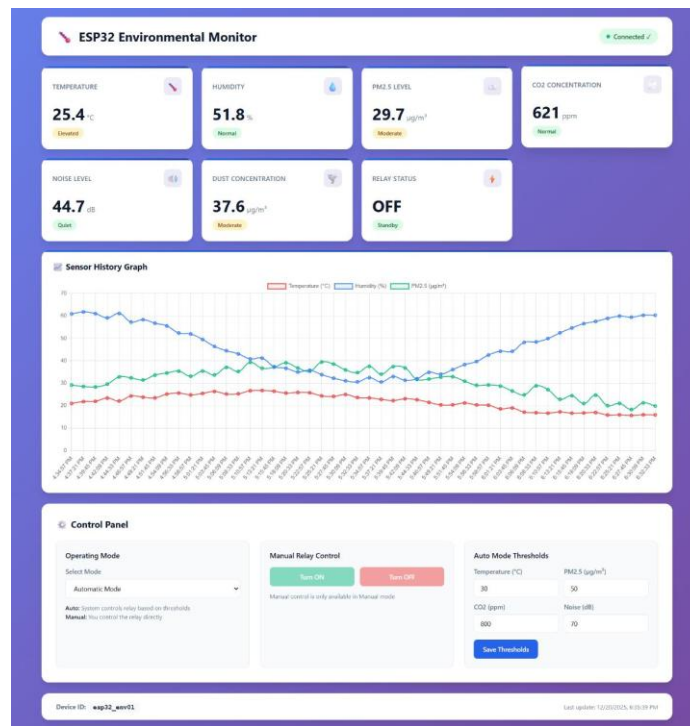
ESP32 được sử dụng để:

- Đọc dữ liệu cảm biến
- Gửi dữ liệu qua WiFi
- Kết nối MQTT

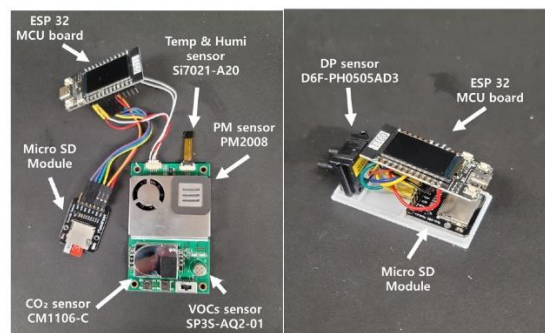
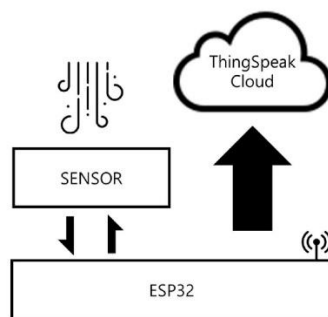
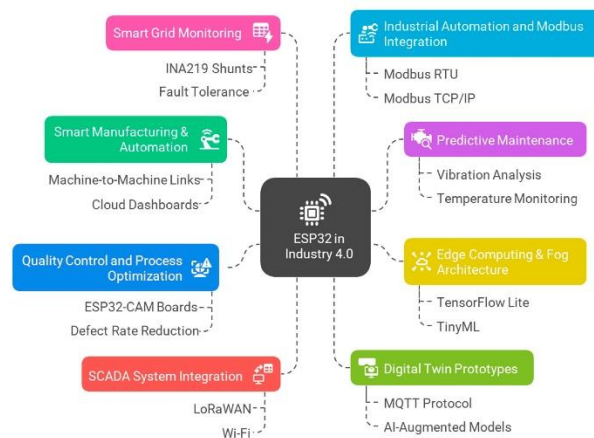
Đặc điểm:

- Tích hợp WiFi/Bluetooth
- Giá thành thấp
- Hỗ trợ IoT tốt
- Tiêu thụ điện năng thấp





ESP32 Applications in Industry 4.0



V.4.2. Cảm biến bụi PMS5003

Cảm biến PMS5003 dùng để đo:

- PM1.0
- PM2.5
- PM10

Thông số:

Thông số	Giá trị
Điện áp	5V
Giao tiếp	UART
Độ chính xác	Cao
Ứng dụng	Giám sát bụi



V.4.3. Cảm biến khí MQ135

MQ135 dùng để phát hiện:

- CO₂
- NH₃
- Benzene
- Khói độc

Ứng dụng:

- Giám sát khí độc
- Phát hiện rò rỉ khí

V.4.4. Raspberry Pi / Jetson Nano

Edge Device thực hiện:

- AI inference
- Phân tích dữ liệu
- Điều khiển hệ thống

So sánh thiết bị

Thiết bị	Ưu điểm	Hạn chế
Raspberry Pi	Giá rẻ	AI hạn chế
Jetson Nano	GPU mạnh	Tiêu thụ điện cao

V.5. Thiết kế phần mềm hệ thống

V.5.1. Kiến trúc phần mềm

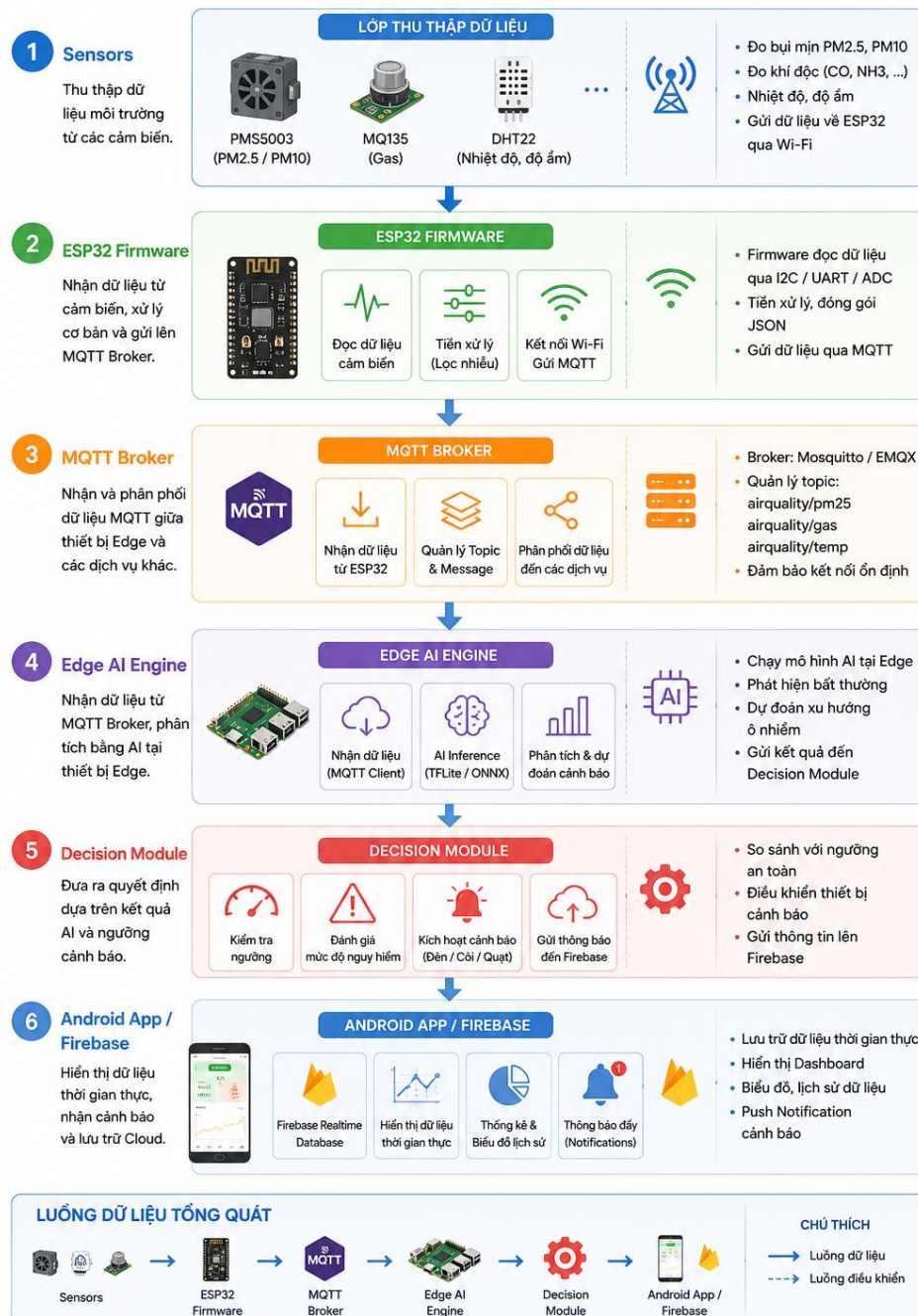
Hệ thống phần mềm gồm:

- Firmware ESP32
- MQTT Broker
- Edge AI Engine
- Android Application
- Firebase Cloud

Sơ đồ phần mềm

KIẾN TRÚC PHẦN MỀM HỆ THỐNG EDGE AI GIÁM SÁT KHÔNG KHÍ

Luồng dữ liệu và xử lý trong hệ thống AIoT



V.5.2. Giao thức MQTT

MQTT được sử dụng để:

- Gửi dữ liệu sensor

- Đồng bộ thời gian thực
- Giảm băng thông

Topic ví dụ:

factory/airquality/pm25

factory/airquality/co

factory/alert/fire

V.5.3. Firebase Realtime Database

Google Firebase hỗ trợ:

- Lưu dữ liệu thời gian thực
- Đồng bộ Android App
- Notification

V.6. Mô hình Edge AI trong hệ thống

V.6.1. Dữ liệu đầu vào

Dữ liệu gồm:

- PM2.5
- PM10
- CO
- Nhiệt độ
- Độ ẩm

V.6.2. Tiền xử lý dữ liệu

Các bước:

1. Lọc nhiễu
2. Chuẩn hóa dữ liệu
3. Trích xuất đặc trưng

V.6.3. AI Model

Các mô hình có thể sử dụng:

- Random Forest
- LSTM
- CNN
- TinyML

Ví dụ:

Mô hình AI dự đoán:

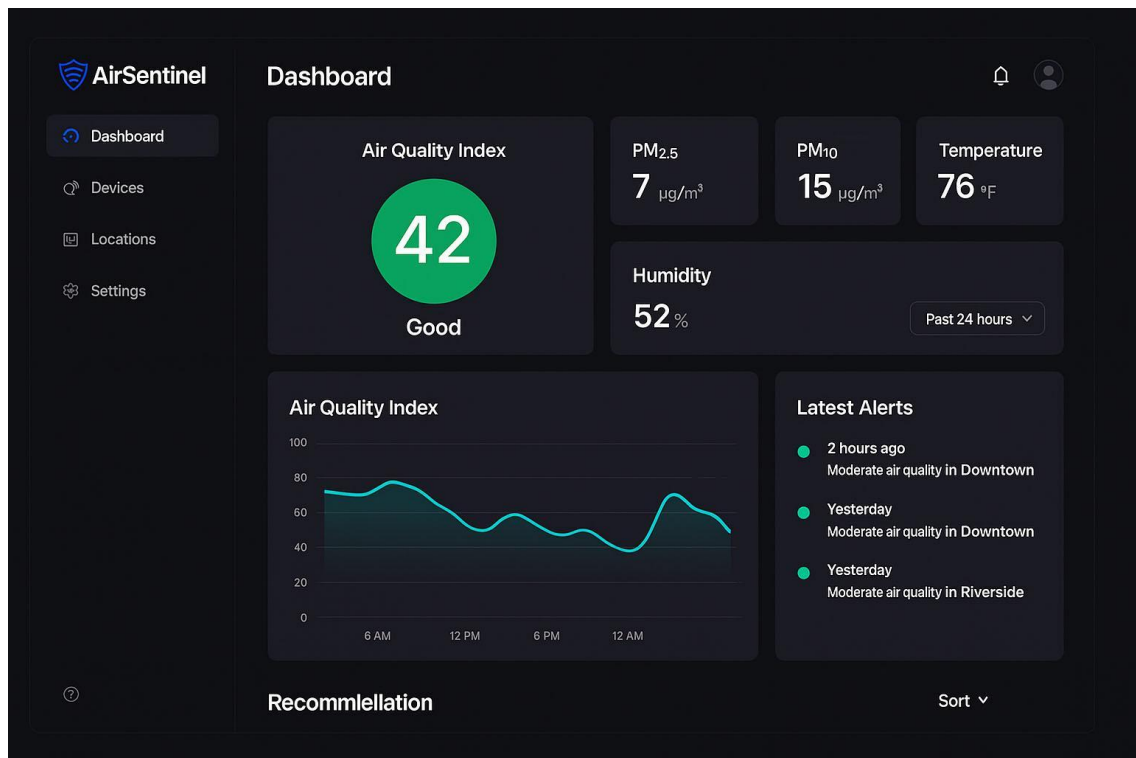
- Nguy cơ ô nhiễm cao
- Xu hướng tăng bụi
- Khả năng cháy nổ

V.6.4. AI Inference tại Edge

Edge Device thực hiện:

- Chạy TensorFlow Lite

- Phân tích thời gian thực
- Ra quyết định cục bộ



V.7. Cơ chế cảnh báo tự động

V.7.1. Điều kiện cảnh báo

Ví dụ:

- $PM_{2.5} > 150 \mu g/m^3$
- CO vượt ngưỡng
- Nhiệt độ tăng bất thường

V.7.2. Các hình thức cảnh báo

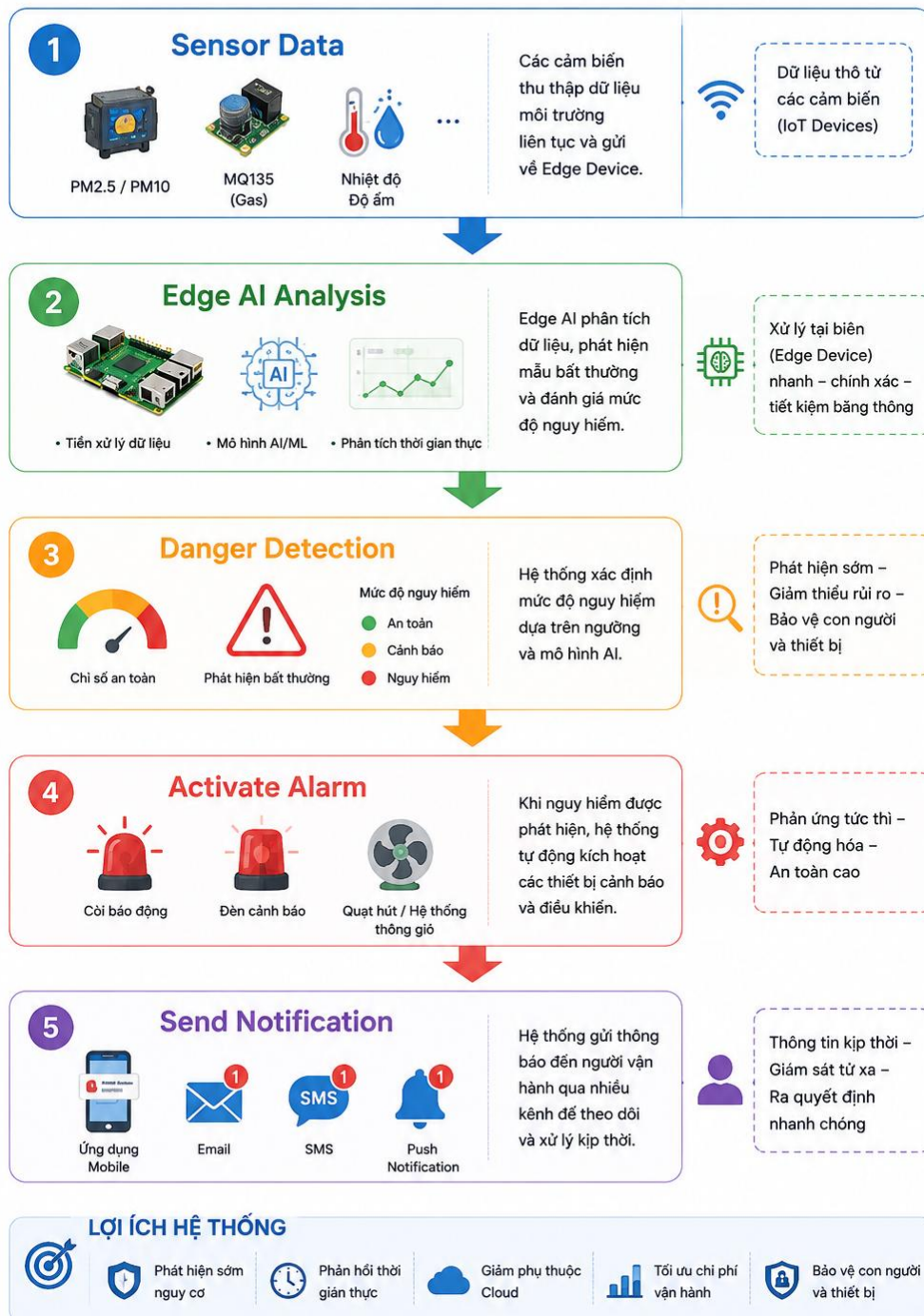
Hệ thống có thể:

- Bật còi
- Bật đèn
- Gửi SMS
- Push Notification Android

Sơ đồ cảnh báo

QUY TRÌNH CẢNH BÁO EDGE AI

Phát hiện – Phân tích – Cảnh báo – Thông báo thời gian thực



Hình 5.2 – Sơ đồ cảnh báo

V.8. Ứng dụng Android giám sát hệ thống

Ứng dụng Android hỗ trợ:

- Hiển thị dữ liệu realtime
- Biểu đồ PM2.5
- Thống kê lịch sử
- Cảnh báo tức thời

Các chức năng:

- Login
- Dashboard
- Alert Screen
- Historical Analytics



V.9. Đánh giá hiệu quả hệ thống

V.9.1. Ưu điểm

Hệ thống:

- Phản hồi nhanh
- Hoạt động offline
- Tiết kiệm băng thông
- Tăng an toàn lao động

V.9.2. Hạn chế

Các hạn chế:

- Edge Device tài nguyên hạn chế
- Cần tối ưu AI model
- Khó triển khai quy mô lớn

V.9.3. So sánh với Cloud AI

Tiêu chí	Edge AI	Cloud AI
Độ trễ	Thấp	Cao
Internet	Không phụ thuộc	Phụ thuộc
Chi phí mạng	Thấp	Cao
Bảo mật	Tốt hơn	Rủi ro cao hơn
AI Model	Nhỏ	Lớn

V.10. Khả năng mở rộng hệ thống

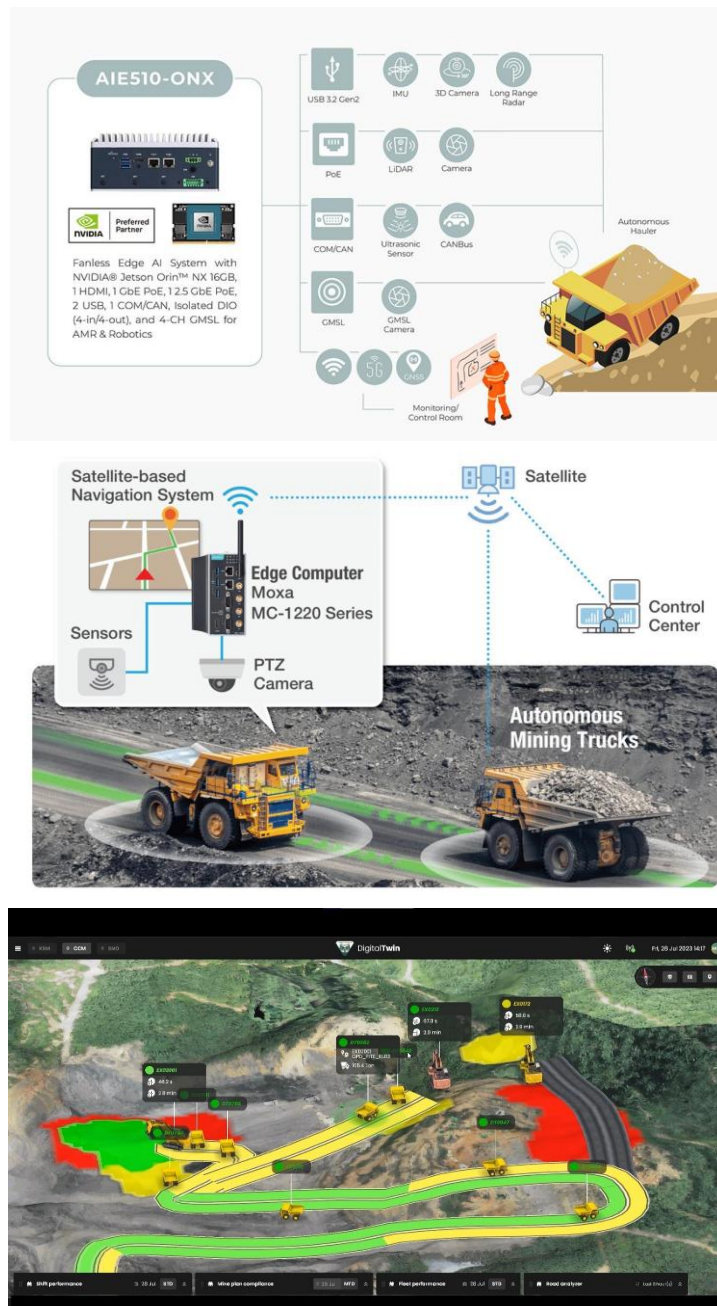
Hệ thống có thể mở rộng:

- Nhiều khu vực mỏ
- Camera AI
- Drone giám sát
- Robot kiểm tra tự động

V.11. Hướng phát triển tương lai

Các hướng phát triển:

- TinyML trên ESP32
- Federated Learning
- Drone AI giám sát bụi
- Edge Generative AI
- Digital Twin cho mỏ thông minh



KẾT LUẬN

Trong bối cảnh chuyển đổi số và sự phát triển mạnh mẽ của Internet vạn vật (IoT), nhu cầu xử lý dữ liệu thời gian thực ngày càng trở nên cấp thiết trong các hệ thống:

- Thành phố thông minh;
- Công nghiệp 4.0;
- Y tế thông minh;
- Giao thông thông minh;
- Giám sát môi trường...

Mô hình xử lý tập trung dựa trên Cloud truyền thống tuy có khả năng tính toán mạnh nhưng vẫn tồn tại nhiều hạn chế:

- Độ trễ cao;
- Tiêu tốn băng thông lớn;
- Phụ thuộc kết nối Internet;
- Rủi ro bảo mật dữ liệu;
- Chi phí lưu trữ và xử lý tăng cao...

Sự ra đời của Edge AI đã mở ra hướng tiếp cận mới cho các hệ thống AIoT hiện đại bằng cách đưa khả năng xử lý trí tuệ nhân tạo xuống gần nguồn sinh dữ liệu hơn. Nhờ đó, hệ thống có thể:

- Phản hồi nhanh hơn;
- Giảm độ trễ;
- Hoạt động ổn định ngay cả khi mất kết nối mạng;
- Tăng cường bảo mật và quyền riêng tư dữ liệu;
- Tối ưu chi phí truyền tải và lưu trữ.

Xu hướng nghiên cứu ứng dụng Edge AI trong IoT:

- TinyML và AI siêu nhẹ trên thiết bị IoT công suất thấp: Sensor thông minh;
- Federated Learning và AI phân tán tại Edge: Smart Healthcare;
- Edge Generative AI và Edge LLM: Robot, Smart Home;
- Edge AI kết hợp 5G/6G và mạng thông minh: Xe tự hành;
- Autonomous Edge Systems và AIoT tự thích nghi: Smart Factory.

TÀI LIỆU THAM KHẢO

- [1] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge Computing: Vision and Challenges,” *Future Generation Computer Systems*, vol. 75, pp. 36–46, Oct. 2017.
- [2] IEEE, “Edge AI: Vision and Challenges,” *IEEE Access*, vol. 10, pp. 9456–9478, 2024.
- [3] Google, “TensorFlow Lite Documentation,” 2025. Available: TensorFlow Lite Documentation
- [4] Google, “Edge TPU Documentation,” 2025. Available: Google Coral Edge TPU
- [5] NVIDIA, “Jetson Platform for Edge AI and Robotics,” 2025. Available: NVIDIA Jetson Platform
- [6] Raspberry Pi Foundation, “Raspberry Pi Documentation,” 2025. Available: Raspberry Pi Documentation
- [7] Intel, “OpenVINO Toolkit,” 2025. Available: Intel OpenVINO Toolkit
- [8] Edge Impulse, “TinyML Development Platform,” 2025. Available: Edge Impulse
- [9] Linux Foundation, “LF Edge Whitepaper,” 2024. Available: LF Edge
- [10] International Data Corporation, “Worldwide Edge Spending Guide,” 2025.
- [11] Gartner, “Top Strategic Technology Trends,” 2025.
- [12] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.
- [13] A. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv:1704.04861, 2017.
- [14] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv:1804.02767, 2018.
- [15] H. Brendan McMahan *et al.*, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Proc. Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.
- [16] National Institute of Standards and Technology, “NIST Cybersecurity Framework,” 2024. Available: NIST Cybersecurity Framework
- [17] Microsoft, “Azure IoT Edge Documentation,” 2025. Available: Azure IoT Edge
- [18] Amazon Web Services, “AWS IoT Greengrass,” 2025. Available: AWS IoT Greengrass
- [19] D. Evans, “The Internet of Things: How the Next Evolution of the Internet Is Changing Everything,” Cisco Systems White Paper, 2011.
- [20] P. Mach and Z. Becvar, “Mobile Edge Computing: A Survey on Architecture and Computation Offloading,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.