INTRODUCTION TO EDGE AI

CVPR 2024 Tutorial

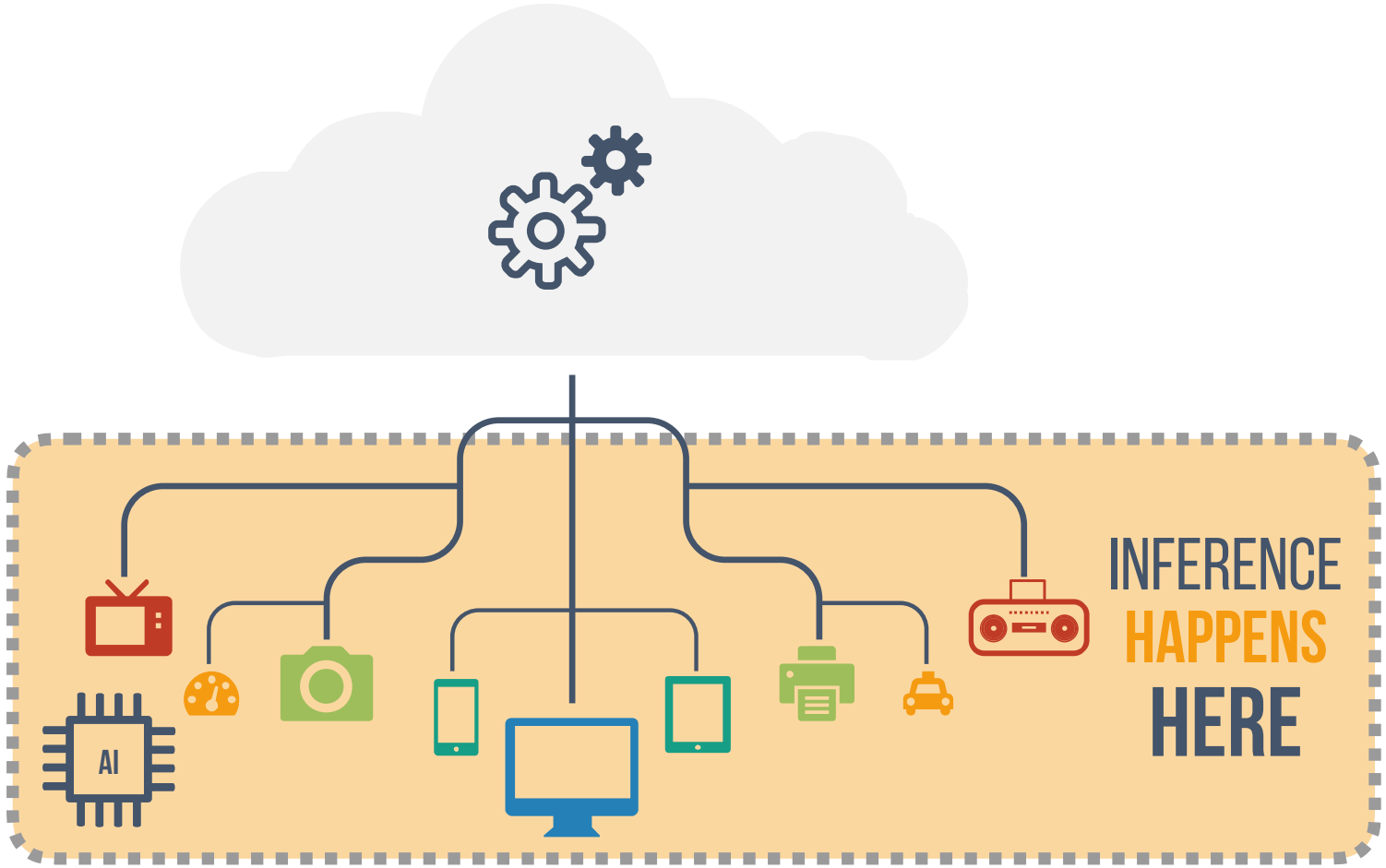
The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024

Seattle, WA, USA

INTRODUCTION:
WHAT
IS EDGE AI?

WHAT IS EDGE AI?

Introduction



1. **Low Latency**
   Local processing significantly reduces response times and improves the performance of real-time applications.

2. **Reduced Bandwidth**
   By processing data on the device itself, Edge AI decreases the volume of data transmitted over the network.

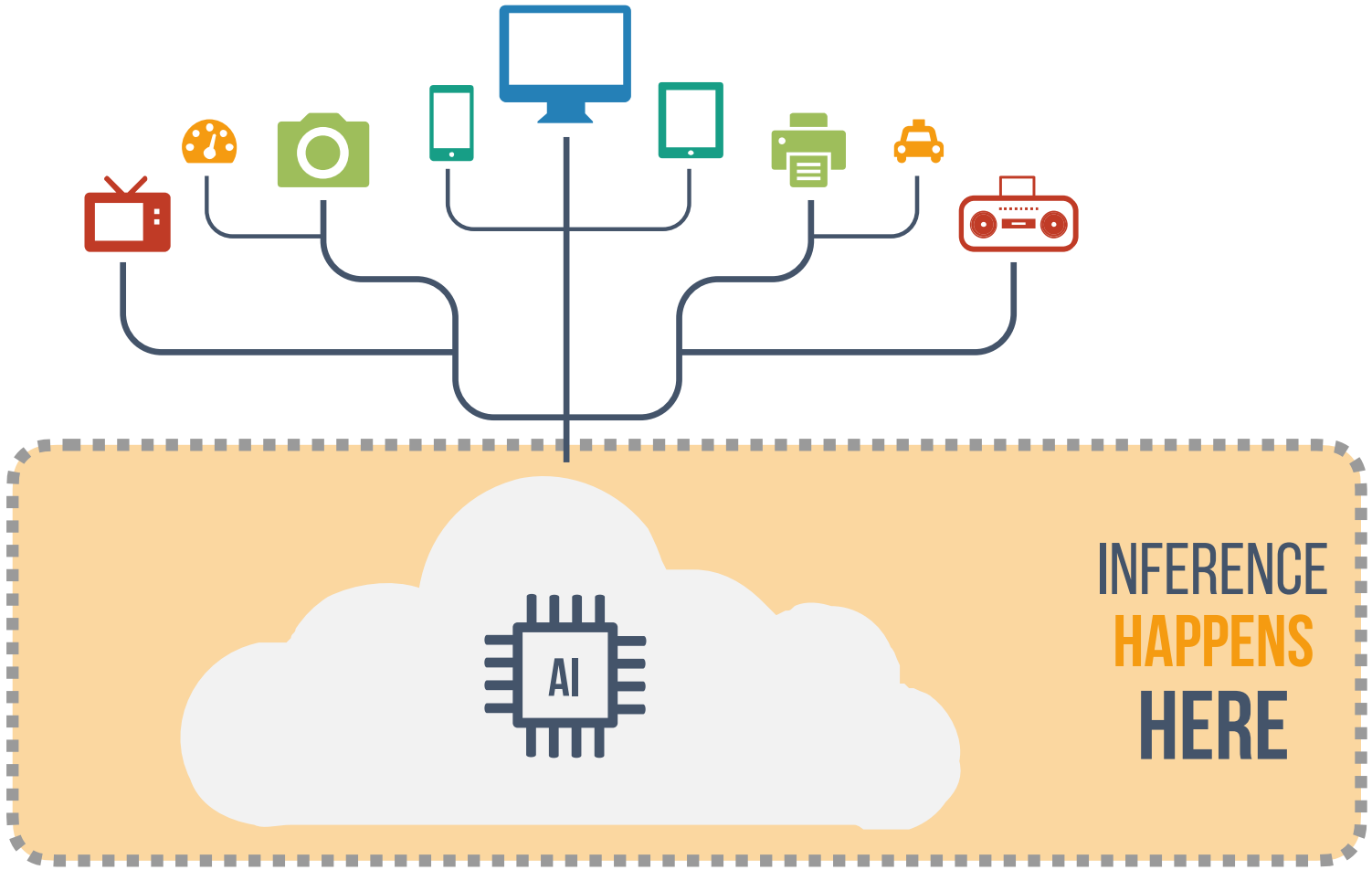
3. **Enhanced Privacy and Security**
   Local data processing means sensitive information does not have to leave the device, enhancing data privacy.

4. **Operational Reliability**
   Edge AI allows devices to operate uninterrupted, independently of the cloud or central servers.

5. **Energy Efficiency**
   Processing data locally can be more energy-efficient than sending data to a cloud for analysis.

WHAT IS CLOUD AI?

Introduction



1. **Scalability**
   Cloud AI systems are highly scalable, allowing for adjustments based on the workload and user demand.

2. **Accessibility**
   Users can access these technologies from anywhere in the world, requiring only an internet connection.

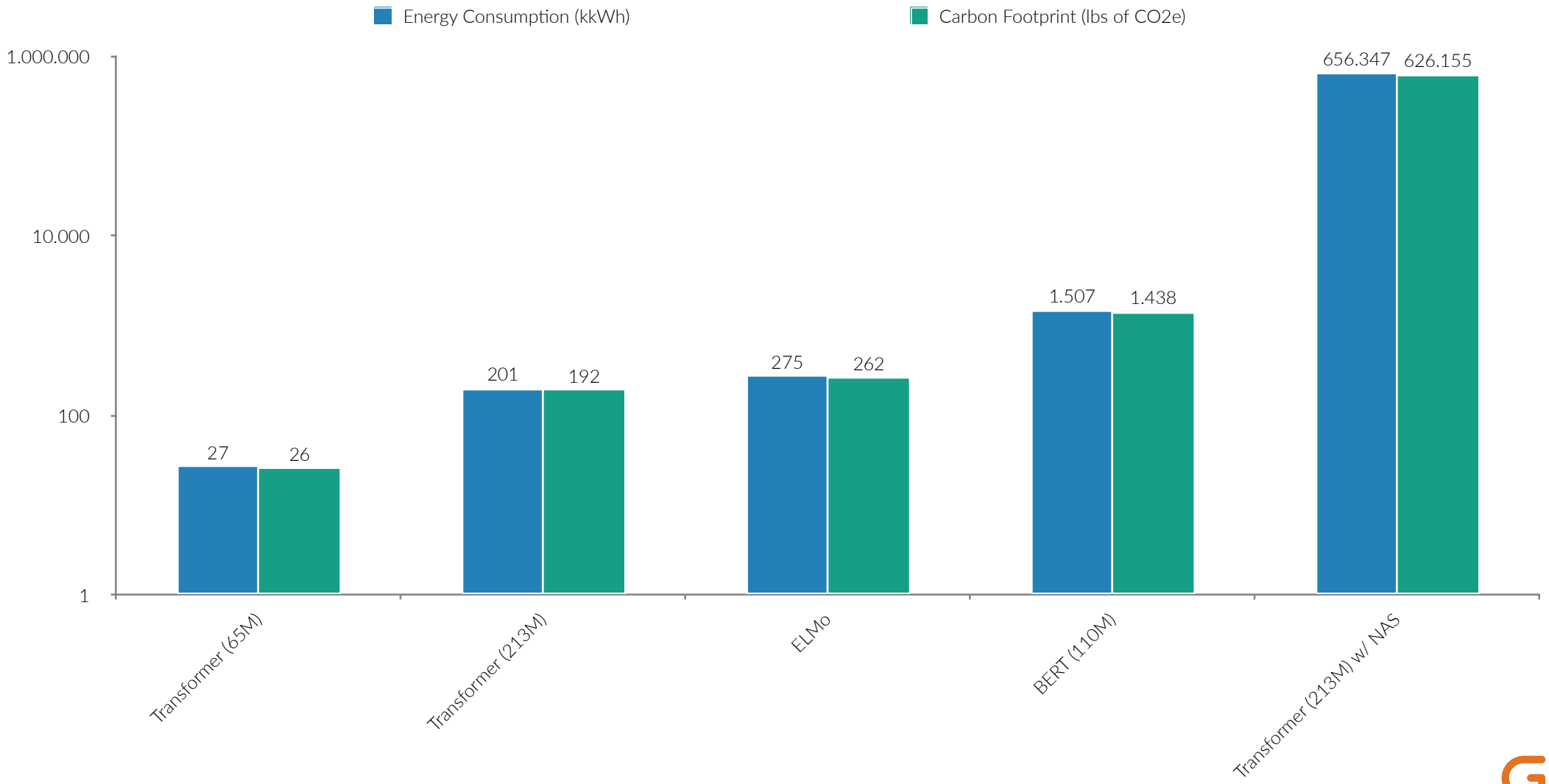
3. **Cost-Effectiveness**
   You can utilize AI tools and computing power on a pay-as-you-go basis, which helps manage costs effectively.

4. **Integration and Collaboration**
   The integration enables seamless data flow and collaboration across different platforms and teams.

5. **Continuous Improvements**
   Cloud AI services are maintained by providers who ensure that the AI models are continuously updated.

WHAT IS THE TRAINING CONSUMPTION?

Training a single AI model can emit as much carbon as five cars in their lifetimes



EDGE AI EXAMPLES

Example in different industries

Tesla Full Self Driving

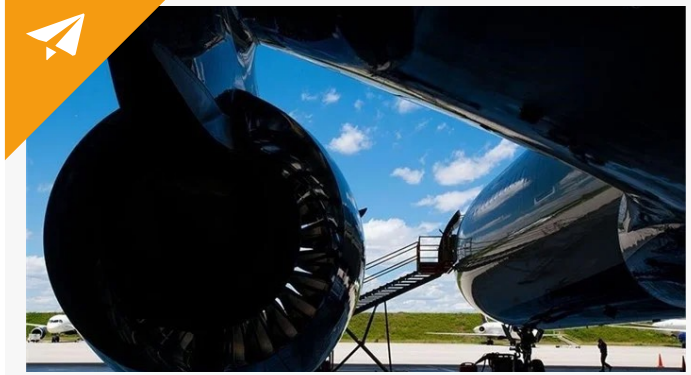
By Tesla

See and Spray

By John Deere

Apple Watch

By Apple

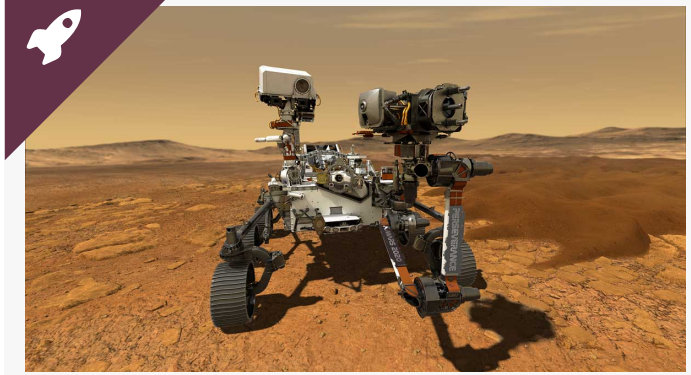


Delta Airlines Predictive Maintenance

By Delta Airlines

Jabra PanaCast 50

By Jabra



Perseverance Mars Rover

By NASA

SECURITY AND PRIVACY

Security & Privacy in Edge AI

As we integrate AI into devices at the edge of our networks, we must adopt robust measures to protect sensitive information and maintain user trust.

Data Encryption

Ensuring data remains encrypted during processing and storage.

Firmware Updates

Protecting devices from unauthorized access and ensuring they run trusted software.

Data Anonymization

Processing data in ways that prevent identification of individuals

Regulatory Compliance

Meeting standards such as GDPR by keeping data processing local

UNVEILING THE PILLARS OF EDGE AI

COMPONENTS OF EDGE AI

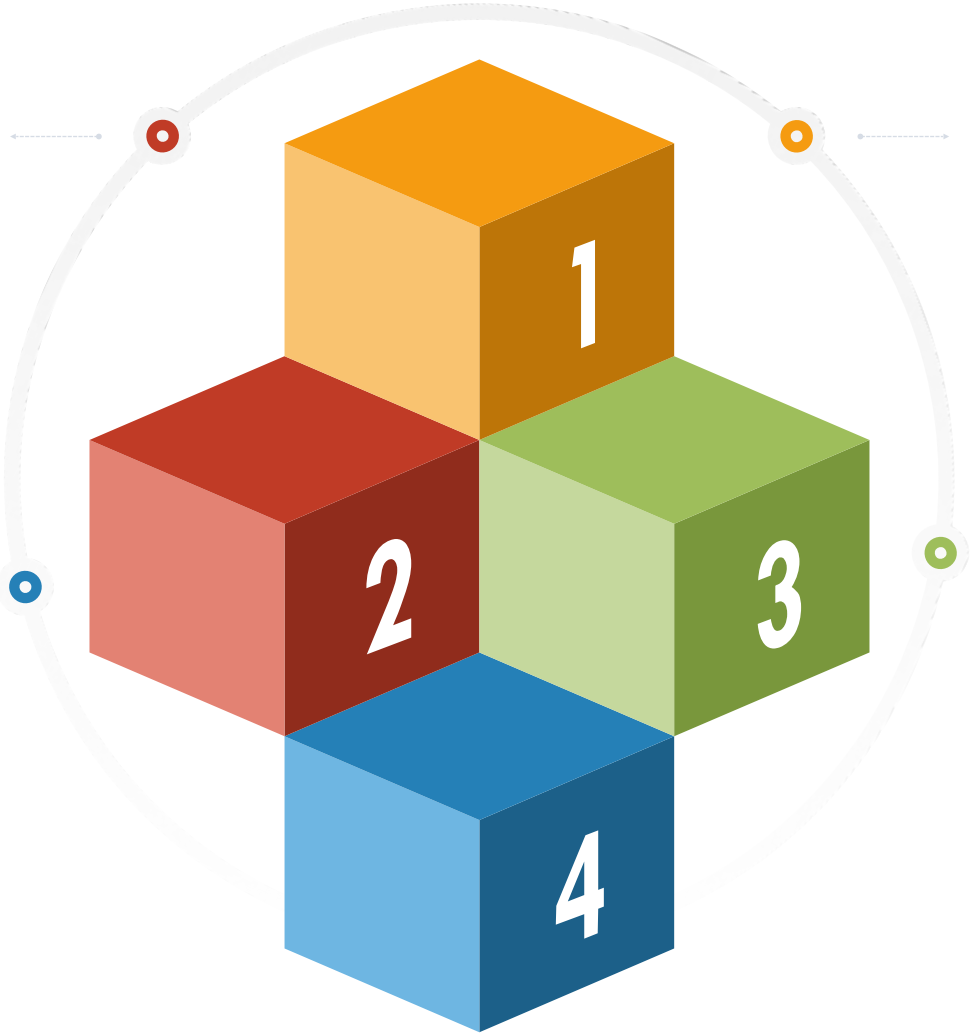
Specialized Hardware and Software

AI Models

Lightweight, optimized models for local inference.

Edge Devices

Incorporate many IoT gadgets, smartphones, cameras, and more.



Software Framework

TensorFlow Lite, PyTorch Mobile, ONNX Runtime.

Hardware Accelerators

Specialized chips (e.g., NVIDIA Jetson, Google Edge TPU).

EDGE AI HARDWARE

Examples of Hardware for Edge AI

Microcontrollers and Microprocessors

Basic computing units for simple AI tasks.

Edge Accelerators

Specialized hardware like NVIDIA Jetson, Google Edge TPU, and Intel Movidius.

Smart Sensors

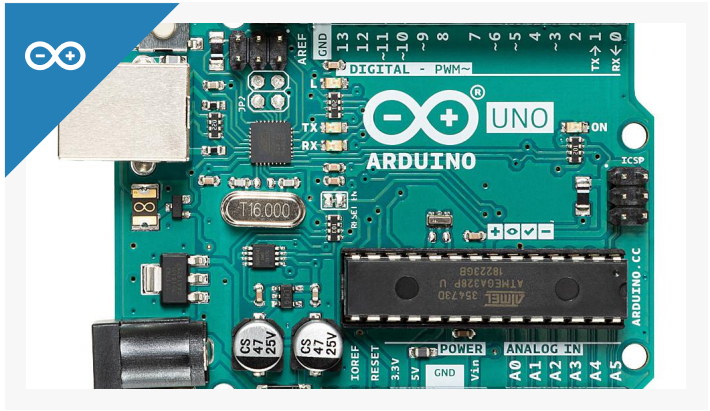
Integrated sensors with built-in AI capabilities for real-time data processing.

Mobile Devices

Smartphones and tablets equipped with AI chips (e.g., Apple's A-series, Qualcomm's Snapdragon).

EDGE AI HARDWARE

Examples of Hardware for Edge AI



Arduino Microcontroller

By arduino.cc

Intel Neural Compute Stick 2

By Intel

BrainChip Akida

By BrainChip

Qualcomm QCS8250

By Qualcomm

NVIDIA Jetson

By NVIDIA

Google EdgeTPU

By Google

EDGE AI SOFTWARE

Frameworks for Edge AI

1 **TensorFlow Lite**
Lightweight version of TensorFlow optimized for mobile and edge devices.

2 **TFLite Micro**
Runs ML models on tiny, low-power devices like microcontrollers.

3 **PyTorch Mobile**
Enables deployment of PyTorch models on mobile platforms.

4 **PyTorch ExecuTorch**
Enables on-device inference capabilities across mobile and edge devices.

5 **Intel OpenVINO**
A toolkit designed to optimize ML and DL models for Intel hardware.

6 **Qualcomm SNPE**
It allows run DL models on Qualcomm Snapdragon mobile platforms.

7 **ONNX Runtime**
Cross-platform, high-performance scoring engine for ONNX models.

8 **Edge Impulse**
A platform for developing, optimizing, and deploying AI models to edge devices.

BRING IT ALL TOGETHER

Workflow for Edge AI Model Deployment

Step 01

MODEL DEVELOPMENT AND TRAINING (PYTORCH, TF, KERAS, JAX, ETC)

01

Trained model weights & model computational graph

Step 02

MODEL OPTIMIZATION (PRUNING, QUANTIZATION, KD, ETC)

02

Optimized model weight & model computational graph

Step 03

EXPORT TO IR (ONNX)

03

IR model weight & computational graph with optimizations.

Step 04

MODEL COMPILATION & BINARY GENERATION (.BIN, .DLC, .XML, .BLOB, ETC)

04

Compiled binary file ready for deployment on edge devices.

Step 05

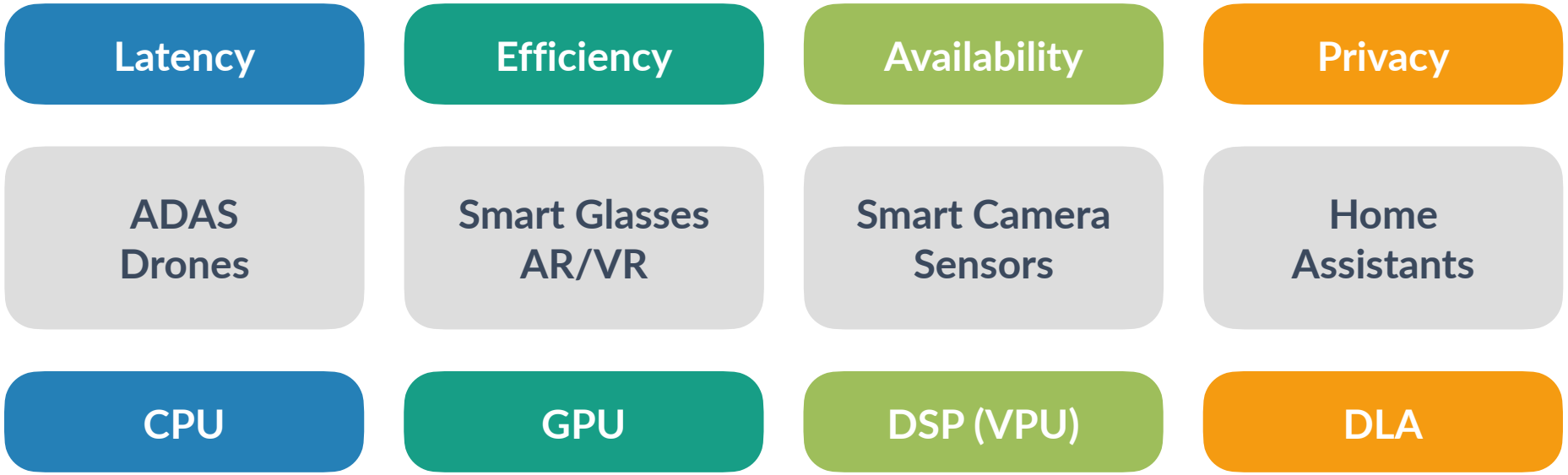
MODEL INTEGRATION & INFERENCE (.C, .H, .CPP, .HPP)

05

NAVIGATING THE CHALLENGES AND PRIVACY LANDSCAPE OF EDGE AI

CHALLENGES IN EDGE AI DEPLOYMENT

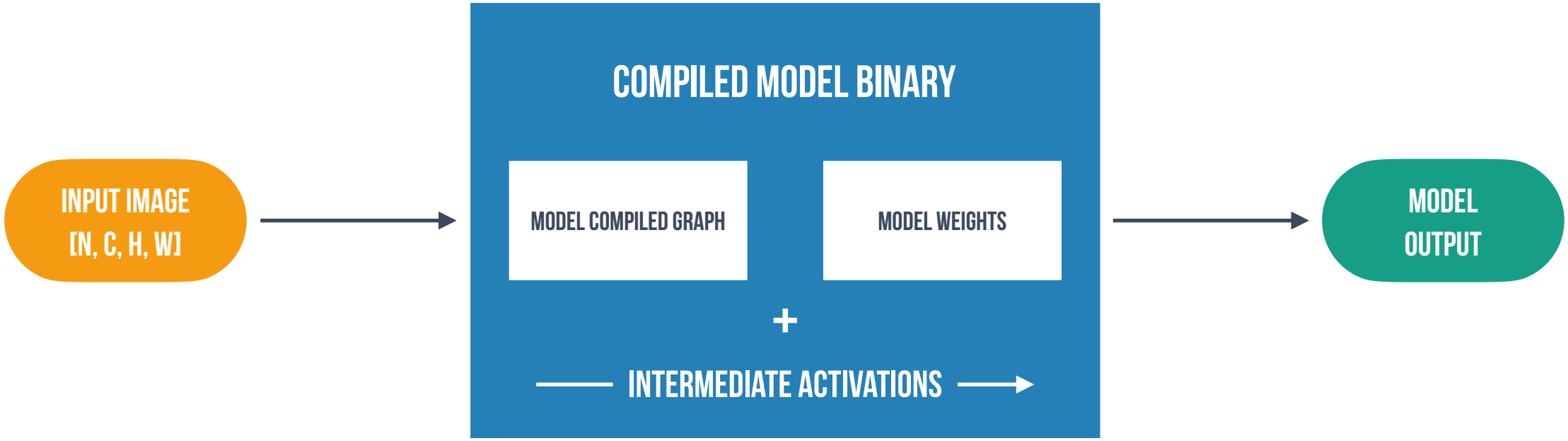
Diagram



INTRODUCTION TO MODEL DEPLOYMENT FOR EDGE AI

MODEL INFERENCE MEMORY BANDWIDTH

Per Frame Model Inference Memory Bandwidth Components



THE ROOFLINE MODEL

Operational Intensity (ops/byte)

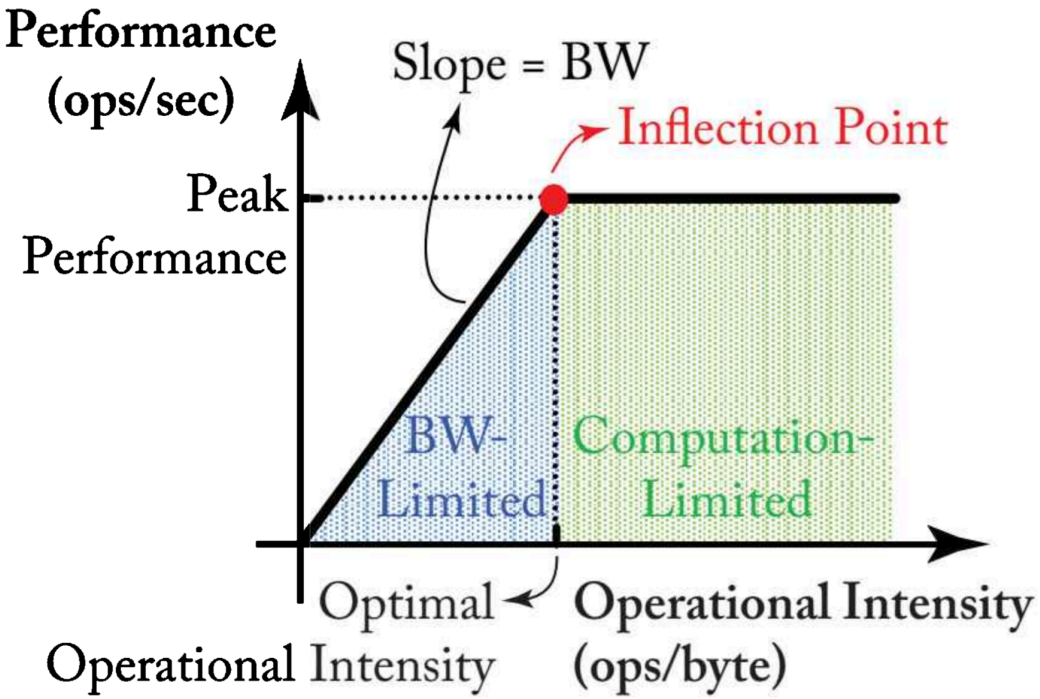
THE ROOFLINE MODEL IS A GRAPHICAL REPRESENTATION TO ILLUSTRATE AN ARCHITECTURE'S PERFORMANCE ACROSS DIFFERENT LEVELS OF OPERATIONAL INTENSITY.

Operational Intensity

How computation-heavy an operation is relative to data movement.

Higher Operational Intensity

More computations are performed for every byte fetched from memory.



THE ROOFLINE MODEL

THE ROOFLINE MODEL

Performance (ops/sec)

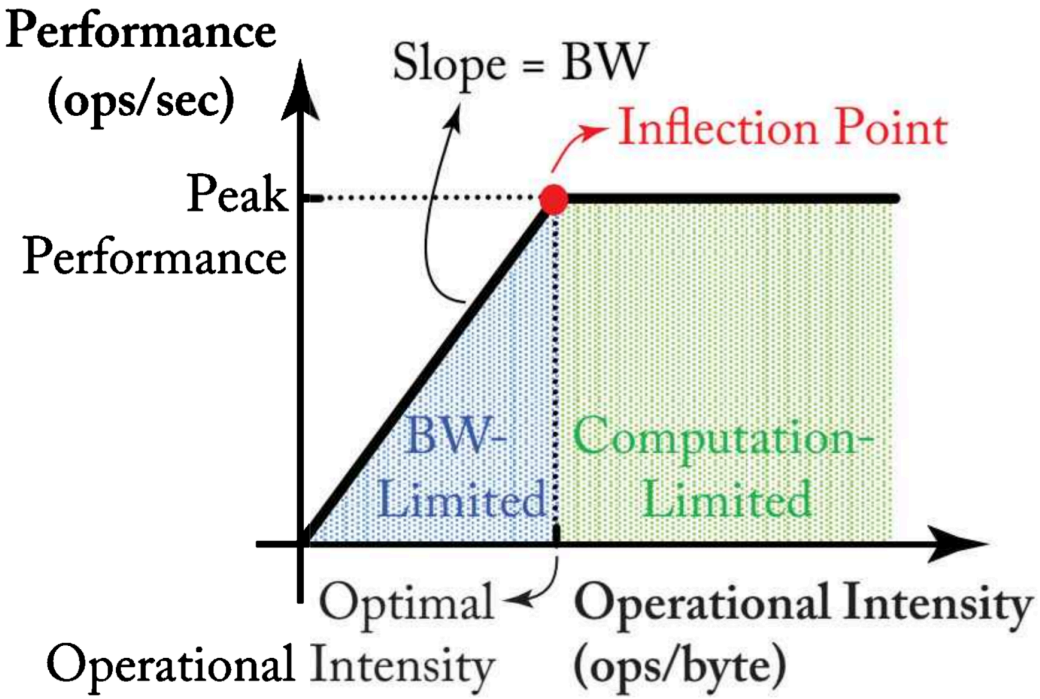
THE ROOFLINE MODEL IS A GRAPHICAL REPRESENTATION TO ILLUSTRATE AN ARCHITECTURE'S PERFORMANCE ACROSS DIFFERENT LEVELS OF OPERATIONAL INTENSITY.

Performance

Represents peak performance of the hardware.

Peak

Maximum number of operations your hardware can handle per second.



THE ROOFLINE MODEL

QUESTIONS & ANSWERS

THANK YOU!