Edge Intelligence: Concepts, architectures, applications and future directions

Dr. Rajiv Misra

Professor, **Dept. of Computer Science & Engg. Indian Institute of Technology Patna** **rajivm@iitp.ac.in**

INTRODUCTION

With the increase in the number of devices connected to a Cloud system, Cloud Computing suffers from certain limitations regarding the high bandwidth requirements to transmit data to the centralized Cloud architecture, high computational power to process the data, and therefore high latency of data processing.

To overcome these limitations a low latency network to process and return data faster to the request sender is required.

Different solutions have been devised to solve these problems. The main approaches are Edge Computing, and Edge AI.

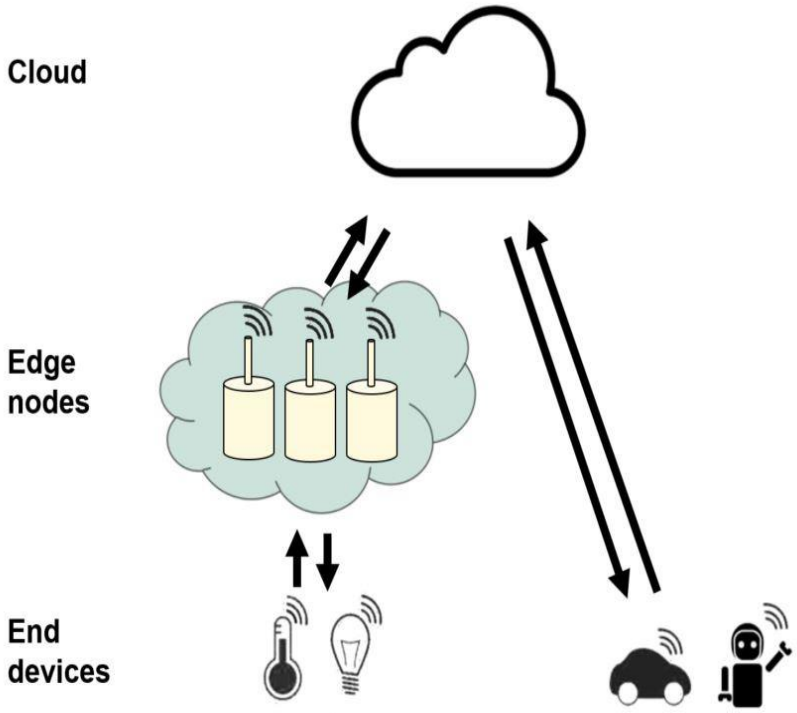
In Edge AI, end devices also act as a data producers, and ML algorithms are used to process, acquire, summarize or transform this data, in the network edge nodes, and sometimes in the end devices themselves.

Edge Computing

Edge Computing is aimed at reducing Cloud workload to process device data, by means of performing some preprocessing and/or computing tasks at the network edge.

Suitable for Big Data analytics and scenarios where a real-time response is required for the user, or where the end device application has time criticality constraints.

Perform some data preprocessing and/or computing tasks at the network edge, instead of relying on external service providers as in Cloud computing.



Edge Computing

Benefits of Edge computing, including ML capabilities at the network edge provides several advantages:

(a) Executing the raw data preprocessing at the network edge can reduce data dimensionality to extract contextual knowledge that can be transferred to the Cloud for higher-level analyses to reduce bandwidth requirements;

(b) it also enables the possibility to perform sensor data fusion at the network edge, as in surveillance scenarios, reducing the workload in Cloud platforms; and

(c) using the computing capabilities of devices and edge nodes helps to accelerate the response of distributed ML algorithms as in Big Data scenarios.

Edge Computing

- An indirect benefit of the Edge Computing approach is the abstraction of the end device. When heterogeneous end devices are used to gather information that complements each other, an adaptation layer is required to transform all the data into a common structure.
- In Edge Computing, a abstraction layer is included inside each end device where relevant information is shared exclusively for the fusing process. This strategy alleviates the computational power required to fuse the raw data in Cloud servers.

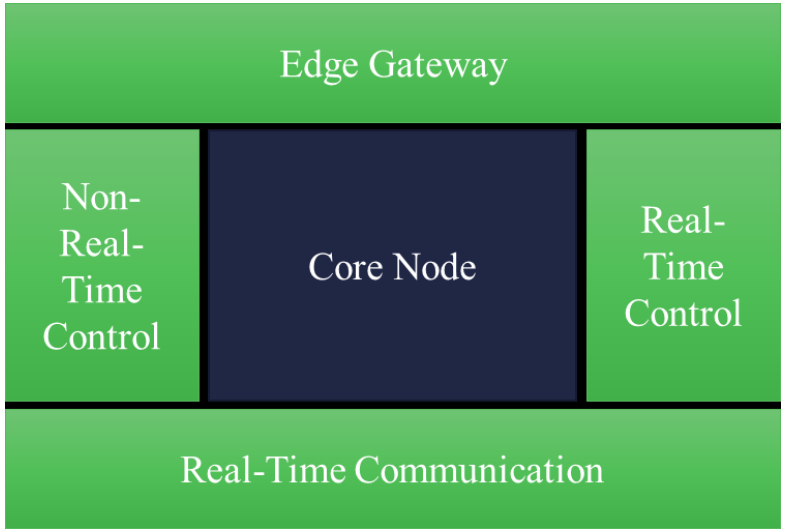
Application scenarios of Edge Computing

Some examples of the applications where the concept of Edge Computing are used:

- Autonomous driving
- Security solutions
- IoT applications
- Location services
- Network functions

Edge Computing Architecture

Fig. shows a modular architecture for Edge Computing devices where real-time is ensured at the same time. Non-real-time and real-time control units perform a control of communication protocols as well as memory allowance



A central core would run an operating system and track the resources of the device to decide where to execute each task of the global process

Besides this, the device should be accessible for updating purposes in order to enable new functionalities and improve current ones.

Scalability is also a key factor in Edge Computing. Therefore, devices should be aware of neighbor devices as well as their functionalities to establish efficient network resource management.

The communication protocols of the device must enable communications not only with other edge devices but also with other devices at the network edge.

Lambda-CoAP Architecture

Abstracts the heterogeneous devices at the network edge while working on real time.

This architecture is composed of different modules that provide the edge devices with functionalities for processing, analyzing and consuming data.

It splits the input data into three different layers in order to reduce the latency.

- **Real time layer**. This layer processes data that needs to be computed on real-time.
- **Batch layer**. This layer preprocesses data to generate batches using the processing of historical data.
- **Serving layer**. This layer displays and communicate the computed results. At the same time, by using this layer it is also possible to access all generated data.

Software Frameworks for Edge Computing

- Apache Kafka
- FAR-Edge RA
- F-Cooper
- Macchina.io
- OGEMA
- CRESCO
- Edge-computing-embedded-platform (ECE platform)
- SA Framework
- Cisco Fog Director
- Crosser
- EdgeX Foundry
- Edgent
- Edge Computing RA 2.0 (EC RA)
- Industrial Internet Consortium RA (IIC RA)
- PiCasso

EDGE INTELLIGENCE

- Edge Artificial Intelligence (Edge AI) or Edge Intelligence can be understood as the confluence of Artificial Intelligence and Edge Computing.

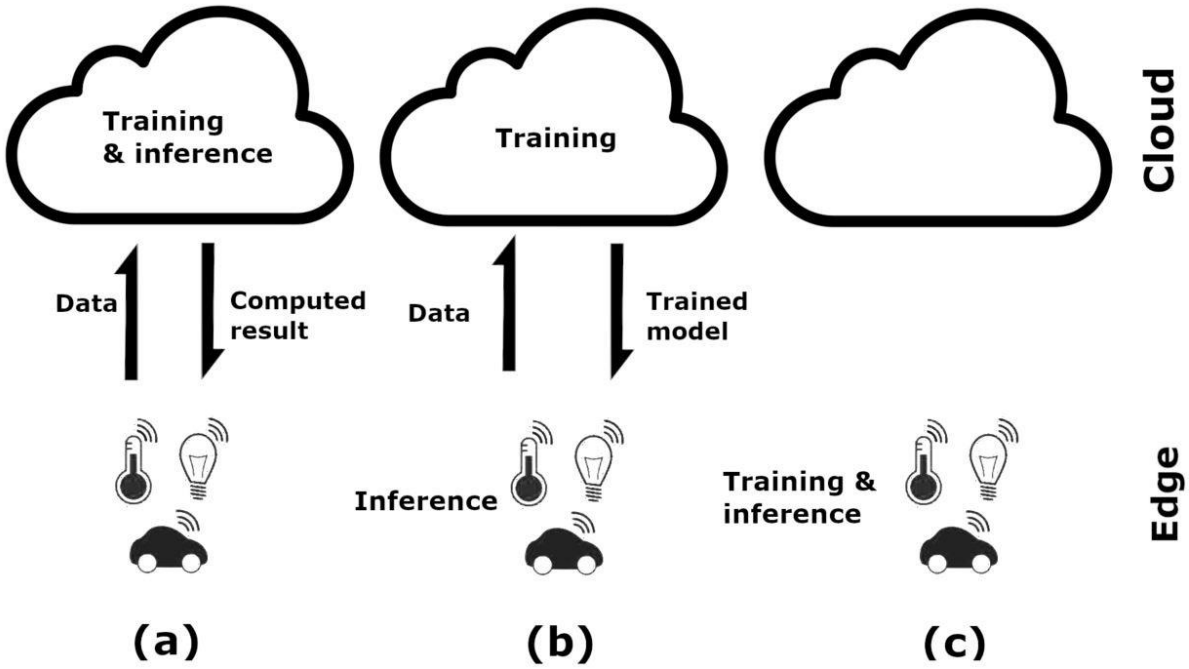
- The goal of this technology is to bring AI capabilities to the network edge.

- Therefore, Edge AI end devices must have enough computational power to run AI inference algorithms that process data with time constraints, and, in specific applications, learning algorithms as well.

Deep Learning at the Network Edge

To fit the constraints of of time and security in applications such as autonomous cars or health care, new approaches for DL at the network edge are emerging:

(a) Traditional training and inference in Cloud servers,
(b) training on Cloud server and inference at the Edge and
(c) training and inference at the Edge.



Deep Learning at the Network Edge

New techniques for the training process which reduce the memory footprint in the edge device as well as increase the training speed in low-resource devices are given below:

- Pruning
- Weights quantization
- Federated Learning and Model Partition
- Transfer Learning
- Knowledge Distillation
- Gossip Training

Application scenarios of Edge Intelligence

Following are the popular applications where Edge AI can be implemented

- Computer vision
- Natural language processing
- Internet of Things (IoT)
- Virtual Reality (VR) and Augmented reality (AR)

Hardware for Edge Intelligence development

Edge AI devices currently in the market which accelerate the DL processed by hardware are as follows:

- Neuroshield
- Google Coral Edge TPU
- Nvidia Jetson Nano
- Intel Movidius
- SparkFun Edge
- BeagleBone AI
- SmartEdge Agile and Branium
- ECM3531
- Smart-Edge-CoCaCo
- MediaTek Solution

Software for Edge Intelligence development

- CMSIS-NN kernels
- DeepMon
- TensorFlow Lite
- MXNet
- Cafe2Go
- CoreML3
- DeepCham
- DeepThings
- DeepIoT
- SparseSep
- ML Kit
- AI2GO
- AWS Greengrass
- Foghorn
- Azure IoT Edge
- OpenEI

Thank You!