Week1: Introduction

Edge Computing

C. García garsanca@ucm.es

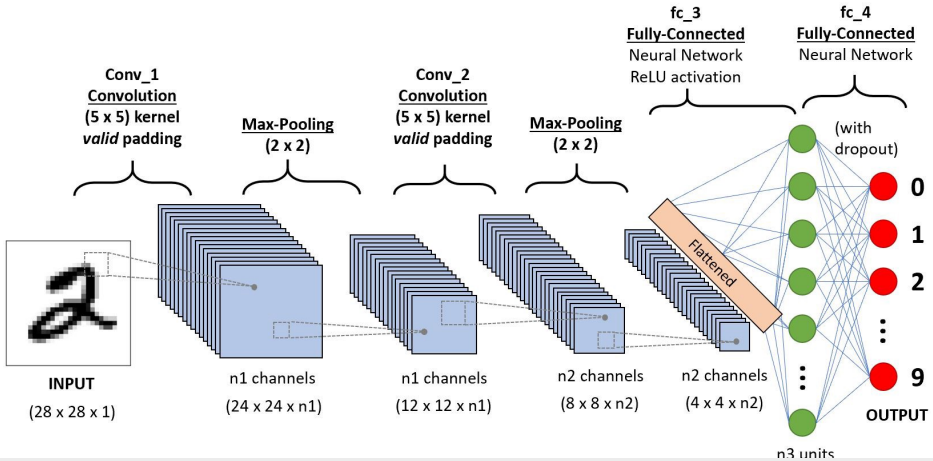
April 18, 2022

- X. Wang et al. "Edge AI, Convergence of Edge Computing and Artificial Intelligence", doi://10.1007/978-981-15-6186-3

Outline

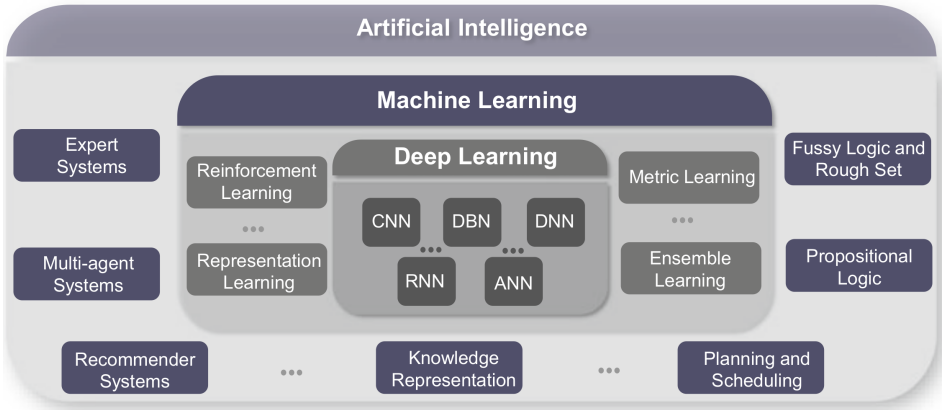
Introduction to AI

- Artificial intelligence (AI) was first proposed at the Dartmouth Conference in 1956
- First approachs: vision that intelligent machines like humans
  - Others image recognition or speech recognition
- Machine learning is one way to implement AI
  - People can train machine learning algorithms to make machines have the ability to learn and reason



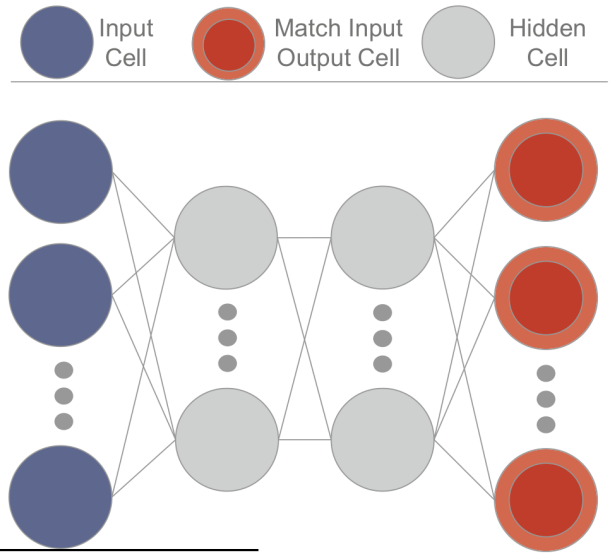
AI, ML and DL

1. **AI** is a relatively broad research area similar to human intelligence
2. **Machine learning** is an important practical way of AI which can make machines have the ability to learn
3. **Deep learning**, as a subset of machine learning, can use neural networks to mimic the connectivity of the human brain to classify datasets and discover correlations between them



Neural Networks in DL

- DL models consist of various types of Deep Neural Networks (DNNs)[1]
  - The most basic neural network architecture is composed of an input layer, a hidden layer, and an output layer.
  - DNNs refers to a sufficient number of hidden layers between the input layer and the output layer

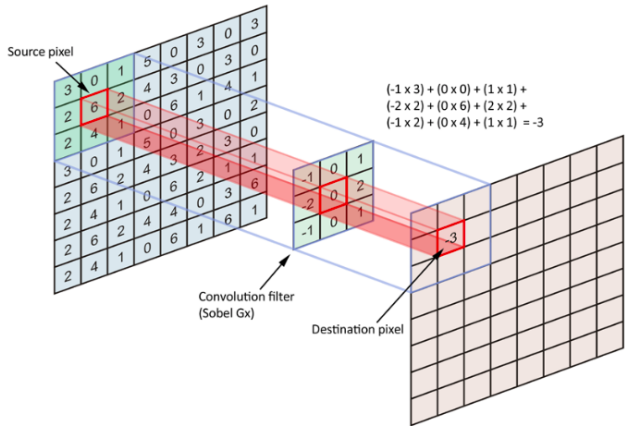


Fully Connected Neural Network

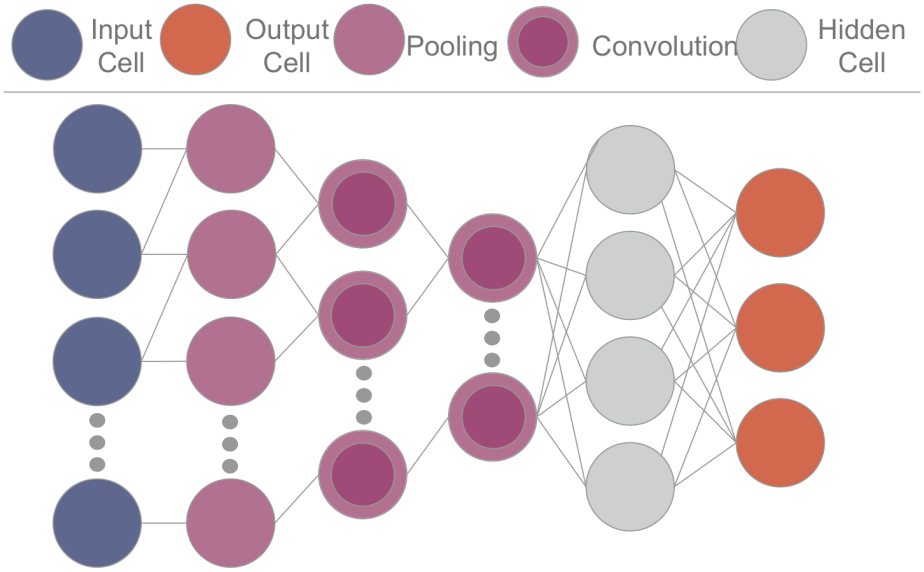
- The output of each layer of FCNN, i.e., Multi-Layer Perceptron (MLP), is fed forward to the next layer
- Between contiguous FCNN layers, the output of a neuron (cell), either the input or hidden cell, is directly passed to and activated by neurons belong to the next layer
- FCNN can be used for feature extraction and function approximation, however, with high complexity, modest performance, and slow convergence.

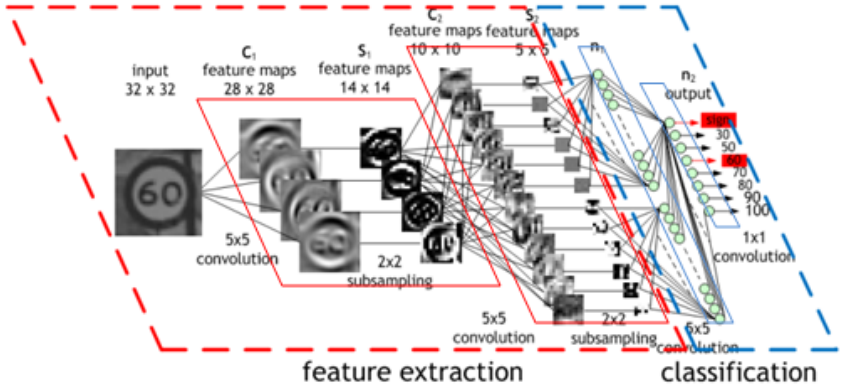
Convolutional Neural Network

- By employing pooling operations and a set of distinct moving filters (convolutional filtering)
  - Convolutional layers: extract image features. Convolutional layers are similar to what we call "filters"
  - Pooling layers: after the operation of the convolutional layer, we get the feature map we want. The pooling layer compresses the feature map



Convolutional Neural Network





Training

- Training DL models in a centralized manner consumes a lot of time and computation resources
- However, DL models have instrinsic parallelism that modern computer can exploit: data parallelism and model parallelism
  - Data paralelism allows to perform a single operation for several data: vector operations at pixel level
  - Model parallelism first splits a large DL model into multiple parts and then feeds data samples for training these segmented models in parallel

DL libraries to optimize

- Development and deployment of DL models rely on the support of various DL libraries:
  1. **TensorFlow**[2] is a relatively high-level framework that can be easily used to design NN structures. Compared to other frameworks, another important feature of TensorFlow is its flexible portability. CPU and GPU support
  2. **Caffe**[3] (Convolutional Architecture for Fast Feature Embedding), the core concept of Caffe is the layer, making input data and perform calculations inside the DL model possible. It is widely used in computer vision such as face recognition, image classification, and target tracking

[2] M. Abadi, P. Barham, et al., TensorFlow: a system for large-scale machine learning

[3] Y. Jia, E. Shelhamer, et al., Caffe: convolutional architecture for fast feature embedding

DL libraries to optimize

- Development and deployment of DL models rely on the support of various DL libraries:
  3. **Theano**[4], the core is a mathematical expression compiler designed to handle large-scale neural network calculations. It compiles various user-defined calculations into efficient low-level codes. However, deploying Theano models is inconvenient and does not support a variety of mobile devices, therefore lacking applications in production environments
  4. **Keras** [5] is a high-level neural network library. Keras is written in pure Python and is based on TensorFlow and Theano. Keras understands deep learning models as an independent sequence or graph

[4] Theano is a Python Library that Allows you to Define, Optimize, and Evaluate Mathematical Expressions Involving Multi-Dimensional Arrays Efficiently

DL libraries to optimize

- Development and deployment of DL models rely on the support of various DL libraries:
  5. **(Py)Torch** [6] support dynamic computation graph, i.e., constructing a graph for each line of code as part of a complete computational graph

[6] A. Paszke, S. Gross, et al., PyTorch: an imperative style, high-performance deep learning library

Intro

- Edge Computing refers to processing the data of the crucial application near production
  - As example, in healthcare, monitoring the older adults at home using the ECG (ElectroCardioGram) or EEG (ElectroEncephaloGram)
  - It is better to perform the data processing at home, and only if something abnormal happens send out the data to the hospital or the doctor
  - Data flow between home and hospital is minimized, which reduces the data failure between two points

Intro

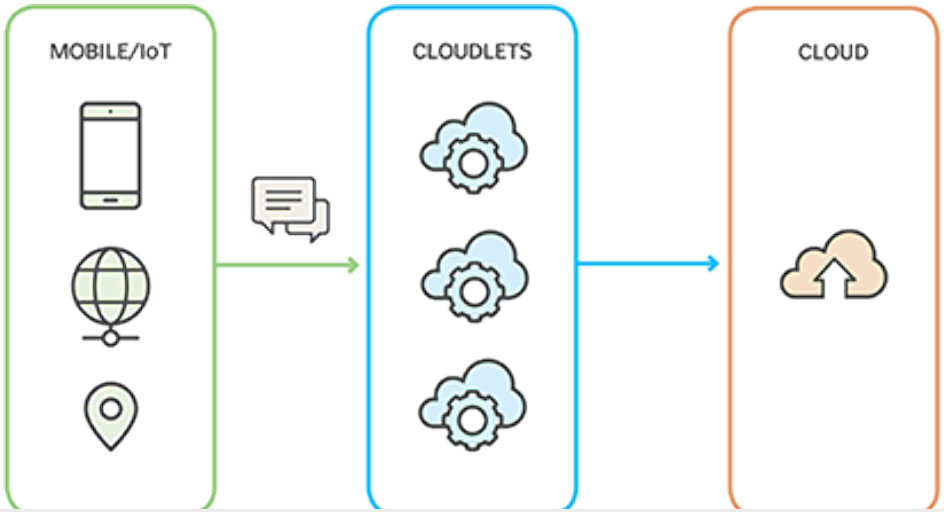
- Electronic devices which are connected to IoT networks can also process the data and exchange the essential information
  - Devices can be drones, robots, cameras, and sensors, etc.

Paradigms

- Various technologies during last years
  - Cloudlet
  - Micro Data Centers
  - Fog Computing
  - Mobile Edge Computing

Cloudlet and Micro Data Centers

- Cloudlet proposed in 2009
  - It is a network architecture combine mobile computing and cloud computing
  - Defines a system and create algorithms (low-latency edge-cloud computing)
- Cloudlet is a data center is a "box": cloud computing closer to the users



Cloudlet and Micro Data Centers

- Cloudlet is a data center is a "box": cloud computing closer to the users
  - Real-time resources to end devices through a WLAN network by running virtual machines on devices
  - It is composed of a set of resource-rich, multi-core computers that have high-speed internet connectivity and high-bandwidth wireless LANs close to devices
- Micro Data Centers was propose by Microsoft
  - It is also designed as a complement of cloud-resources
  - It incorporates flexibility and scalability in terms of capacity and latency by requirements

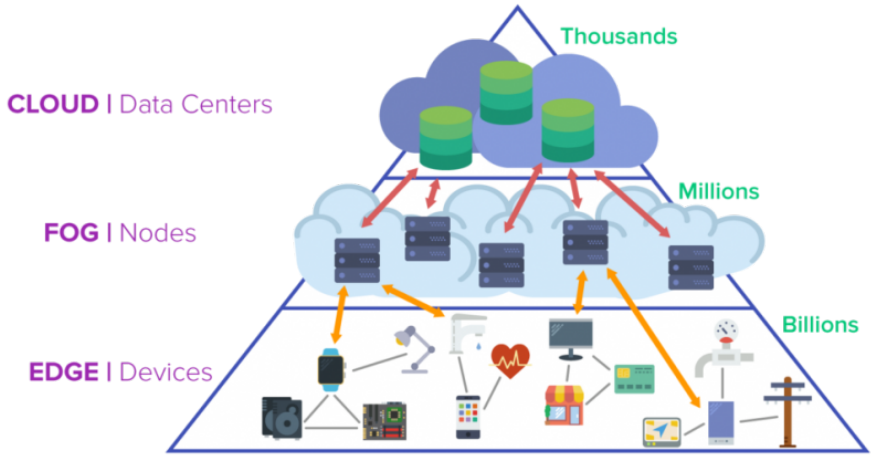
Fog Computing

- Concept proposed by Cisco in 2012 as extension of cloud computing from networks which covers edge
  - OpenFog Consortium[7] which is the main promoter of fog computing
  - It assumes a fully distributed multi-tier cloud computing architecture with billions of devices and large-scale cloud data-centers
- Fog Computing Nodes (FCN) are heterogeneous: routers, set-top boxes, switches, IoT gateways

[7] OpenFog reference architecture for fog computing https://www.openfogconsortium.org/ra/

Fog Computing

- FCN supports devices at different protocol layers and even supports non-IP access technologies to communicate between FCN and end devices
  - Fog abstraction layer is used to hide the heterogeneity of the nodes
  - Providing functions such as data management and communication services between the end device and the cloud



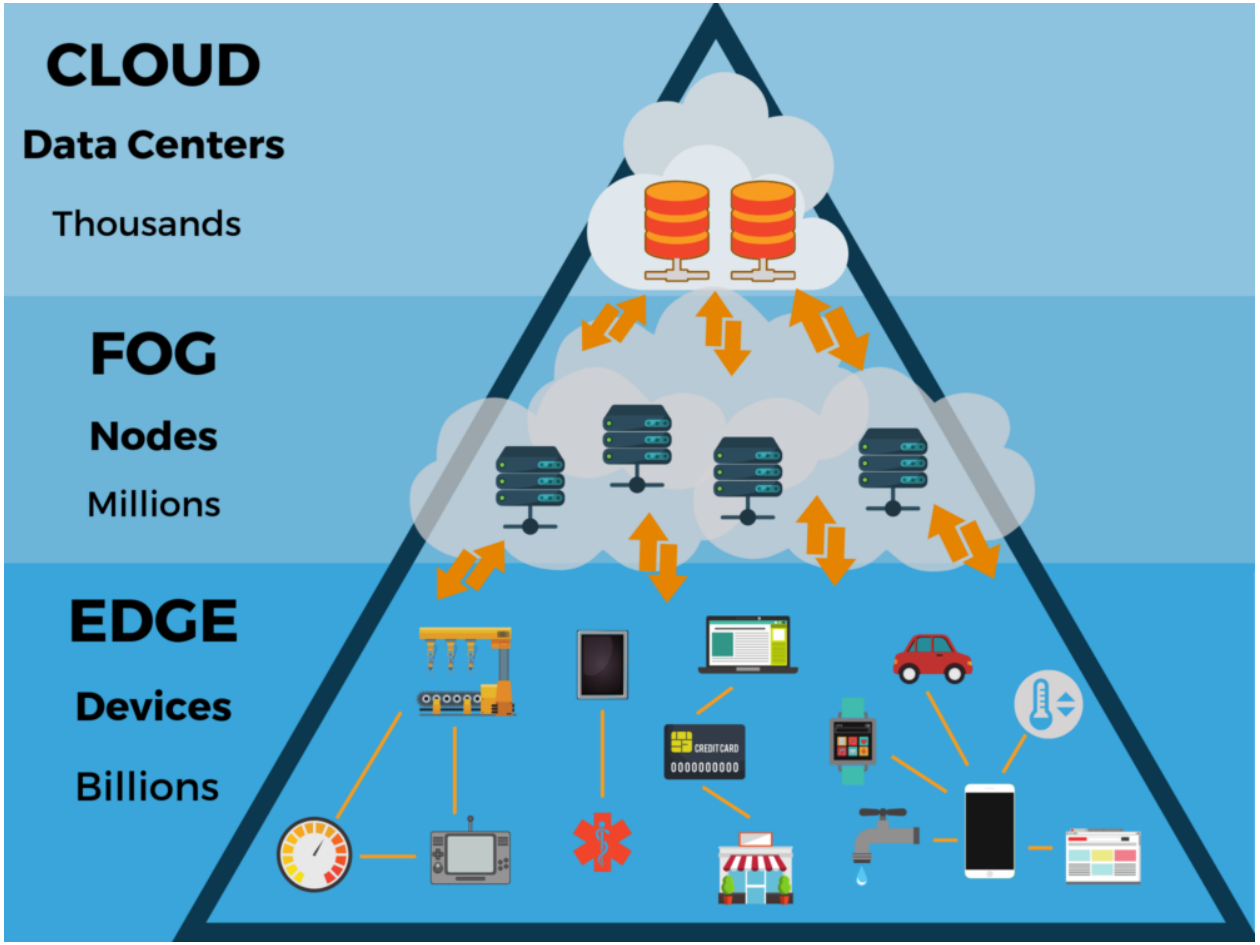
Fog Computing

- Fog computing cannot run on its own without cloud computing
  - It is designed for applications that require real time responding with less latency, such as interactive and IoT applications
- Fog computing **is more focused on IoTs** than Cloudets and Micro Data Centers

Edge Computing

- Standardized by the Mobile Edge Computing Specification Working Group of the European Telecommunications Standards Institute (ETSI) in 2014
- Places computing capabilities and service environments at the edge of cellular networks
  - Designed to provide lower latency, context and location awareness, and higher bandwidth
  - Edge devices are ambiguous in most literature (the boundary between edge nodes and end devices is not clear)
    - Mobile edge devices (including smartphones, smart vehicles, etc.)
    - IoT devices, and the "edge nodes" (edge level) include Cloudlets, Road-Side Units (RSUs), Fog nodes, edge servers

Edge Computing

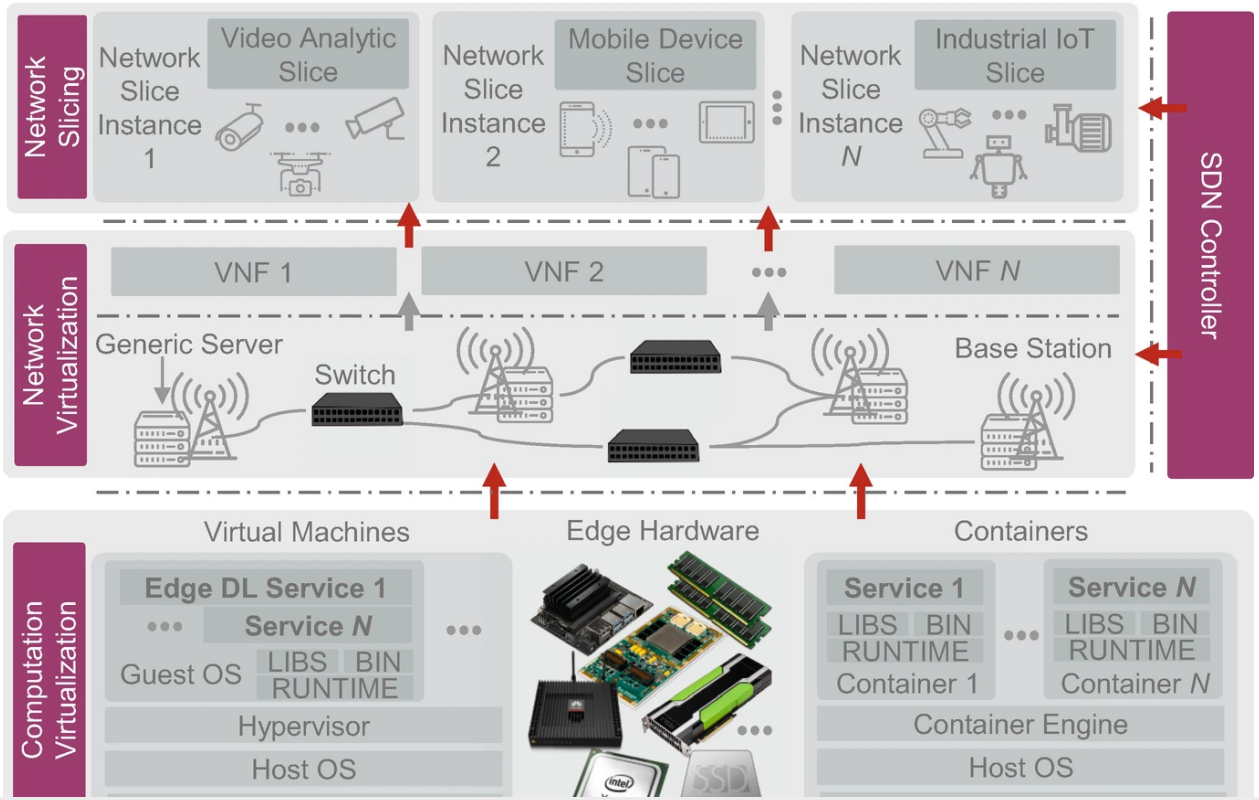


Edge Computing

	Vendor	Name	Features
Integrate Commodities	Intel	Movidius NCS	Prototype on any platform with plug-and-play simplicity
	NVIDIA	Jetson Nano	Easy-to-use platforms that runs in as little as 5 W
	Google	Coral	Local IA platform: efficient, private, fast and private
AI Hw for Edge	HiSilicon	Kirin Series	Independent NPU for DL computation
	NVIDIA	Ampere TPU	Powerful capabilities (but high energy consumption)
	Google	TPU	Stable in terms of performance and power consumption
	Intel	Habana	Lower total cost of ownership (AWS-EC2)
	Samsung	Exynos	GPU and NPU for accelerate inference

Edge Computing

- Architecture
  - Network virtualization: run in software, by separating network functions and services from dedicated network hardware



Scenarios

- Smart park scenario: responsible for the following functions:
  - Massive network connection and management
  - Real-Time data collection
- Video-Surveillance: transition from "seeing" to "understanding"
  - Edge computing can reduce the reliance on cloud computing to improve the storage and computing efficiency
  - Edge node image recognition and video analysis
  - Optimize data storage based on video analysis results
- Industrial Internet of Things: requirements in digitalization
  - Data is heterogeneous from multiple sources: End nodes unify the format through preprocessing
  - Insufficient security protection for key data: Edge computing security mechanisms

Benefits of Edge Computing

- Main benefits of edge computing are:
  - **Lower Latency**: by processing data at a network's edge, data travel is reduced
    - This opens the door to advanced use cases with more complex AI models such as fully autonomous vehicles and augmented reality, which require low latency
  - **Reduced Cost**: using a LAN for data processing means organizations can access higher bandwidth and storage at lower costs compared to cloud computing.
    - Less data needs to be sent to the cloud: reduces costs further.

Benefits of Edge Computing

- Main benefits of edge computing are:
  - **Model Accuracy**: AI relies on high-accuracy models
    - When a network's bandwidth is too low, typically reducing the size of data used for inferencing
    - Reduced image sizes (or skipped frames in video) and reduced sample rates in audio supposes lower accuracy
    - When deployed at the edge, data feedback loops can be used to improve AI model accuracy
  - **Wider Reach**: Internet access is required for traditional cloud computing.
    - Edge computing processes without internet access
  - **Data Sovereignty**: data is processed at the location it is collected
    - Edge computing allows organizations to keep all of their data and compute inside the LAN and company firewall

Use Cases Across Industries

- **Retail**, retailers can boost their agility by:
  - **Reducing shrinkage**: in-store use of cameras and sensors allows to analyze data, stores can identify and prevent instances of errors, waste, damage and theft.
  - **Improving inventory management**: Edge computing applications can use in-store cameras to alert store associates when shelf inventories are low, reducing the impact of stockouts
  - **Streamlining shopping experiences**: fast data processing, retailers can implement voice ordering so shoppers can easily search for items, ask for product information and place online orders using smart speakers or other intelligent mobile devices.

Use Cases Across Industries

- **Smart Cities**: the use of Edge computing to make their spaces more operationally efficient, safe and accessible are some examples:
  - **Reducing traffic congestion**: edge-computing uses computer vision to identify, analyze and optimize traffic.
    - Improve traffic flow, decrease traffic congestion-related costs and minimize the time drivers spend in traffic.
  - **Monitoring beach safety**: image-detection application helps spot dangers at beaches, hazardous ocean conditions, allowing authorities to enact life-saving procedures
  - **Increasing airline and airport operation efficiency**: video analytics could help airlines and airports make better and quicker decisions around capacity, sustainability and safety

Use Cases Across Industries

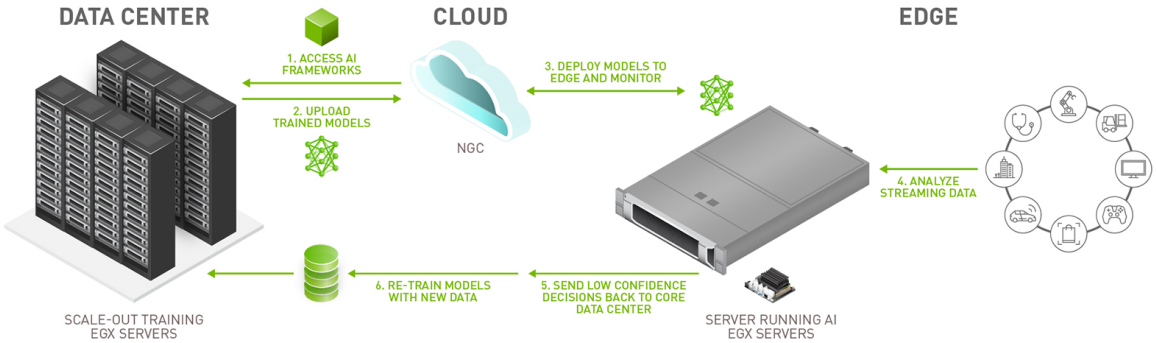
- **Automakers and Manufacturers**: generating sensor data that can be used in a cross-referenced fashion to improve services
  - **Predictive maintenance**: detecting anomalies early and predicting when machines will fail to avoid downtime
  - **Quality control**: detecting defects in products and alerting staff to reduce waste and improve manufacturing efficiency
  - **Worker safety**: cameras and sensors equipped with AI-enabled video analytics to allow manufacturers to identify workers in unsafe conditions and to quickly intervene to prevent accidents

Use Cases Across Industries

- **Healthcare** could be improved by operational efficiency, ensure safety, and provide the highest-quality care experience possible:
  - **Operating rooms**: streaming images and sensors in medical devices could help diagnosis and therapy planning
  - **Hospitals**: patient monitoring, patient screening, conversational AI, heart rate estimation, radiology scanners and more. Human pose estimation can be used to help notify staff when a patient moves or falls out of a hospital bed

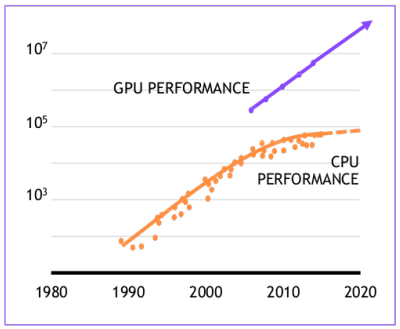
NVIDIA ecosystem

- NVIDIA is a semiconductor company specialized in Graphic Processor Unit building
  - From HPC (supercomputer) to Edge computing

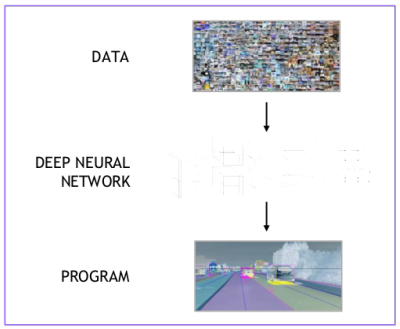


NVIDIA ecosystem

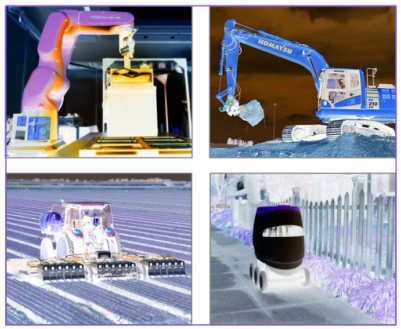
- Brings certified Systems from embedded platforms, AI software and management services to develop AI at the edge



Beyond Moore's Law – 1000x Every 10 Years



Computers Writing Software



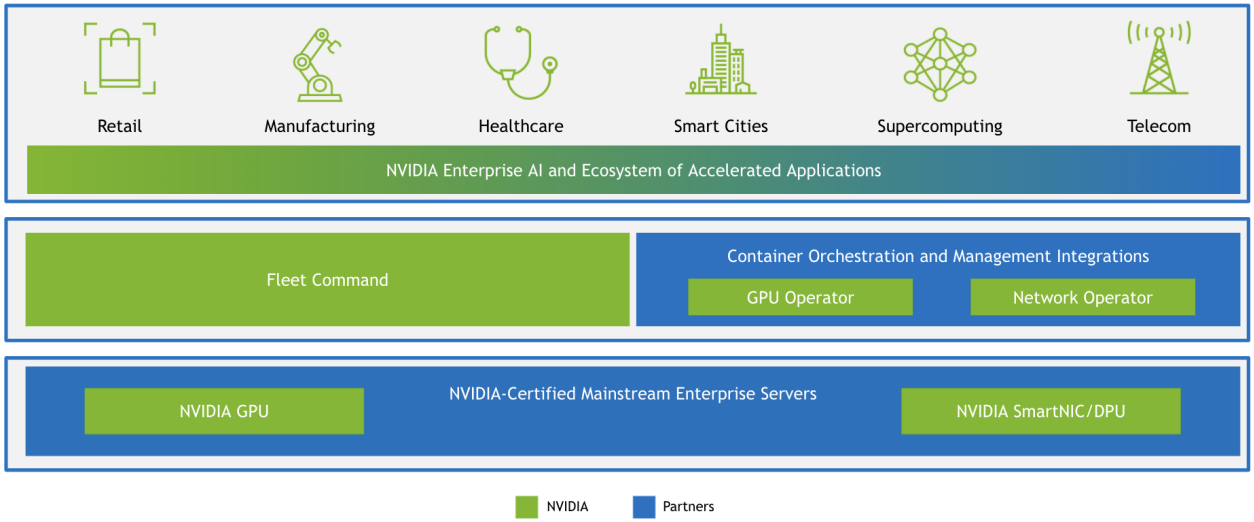
AI-powered Autonomous Machines and Intelligent Systems are Here

NVIDIA ecosystem

- GPU provide a host of software-defined hardware engines for accelerated networking and security
- These hardware engines allow to reach the best performance
  - It is optimally designed for running modern applications
  - Enable CUDA, kubernetes devices plugin for GPUs
  - NVIDIA countainer runtime, automatic labeling

NVIDIA ecosystem

- Accessible and easy to use



NVIDIA Jetson-Nano

- Developer Kit
- AI Computer for less than 99US$

Hardware features

- 128 CUDA Cores | 4 ARM's Core CPU
- Performance Peak: 472 GFLOPs
- Consumption: 5W (GPU) + 10W (CPU)

NVIDIA Jetson-Nano

- Tutorials
- Jetson-Nano Developer Kit
- Getting start