Introduction to Edge Computing and Analytics

Understanding the Edge of Technology

Introduction… to Cloud Computing

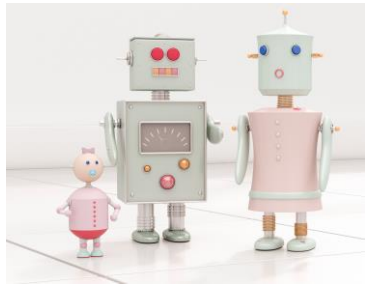
- **Need for Scalable and Flexible Workload Management**
  - Organisations aim to shift from capital expenditures to operational expenses through pay-as-you-go models.
  - Global access and faster deployment of services are in demand.
- **Delivers computing services such as servers, storage, and software over the internet (“the cloud”).**
  - Infrastructure as a Service (IaaS)
  - Platform as a Service (PaaS)
  - Software as a Service (SaaS)
- **Cloud Deployment Models**
  - Public, Private, Hybrid Clouds
- **Key Characteristics**
  - On-demand self-service
  - Broad network access
  - Resource pooling and elasticity
  - Measured service

Inherent Challenges… of Cloud Computing

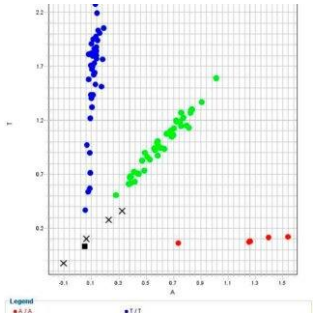
- **Latency and Bandwidth Constraints**
  - Unsuitable for time-sensitive applications
  - Potential increase in data transmission costs
- **Privacy Concerns**
  - Off-premises data storage poses risks
  - Regulatory challenges, e.g. Health data
- **Dependence on Internet Connectivity**
  - Affects service availability
- **Limited Control and Vendor Lock-in**
  - Difficult to customise
  - Complicates migration between providers

Image Credit:
"**Tips: Overcoming Individual & Interaction Challenges**" by **Numrah Khan**, Agile Transformation Expert - CSM®, A-CSM℠, CSPO®, Certified SAFe® 5 Agilist & ICAgile Certified Professional Agile. Source: LinkedIn, URL: https://media.licdn.com/dms/image/v2/D4D12AQHEgNTN3RcqOw/article-cover_image-shrink_720_1280/0/1677410626617?e=1733356800&v=beta&t=o-GzbxaRoClXcpGQqL2Ztkojj8yfhSNgkiISfjJOE10, Accessed 6 October 2024.

A Changing World – Cotton Candy Cloud

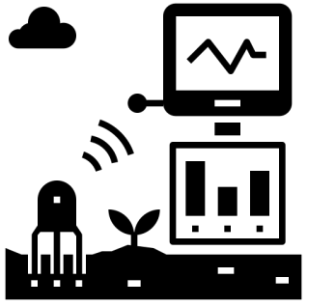


More Devices Online



Machine Leaning in IoT

Devices Generating more Data



Cheaper Sensors

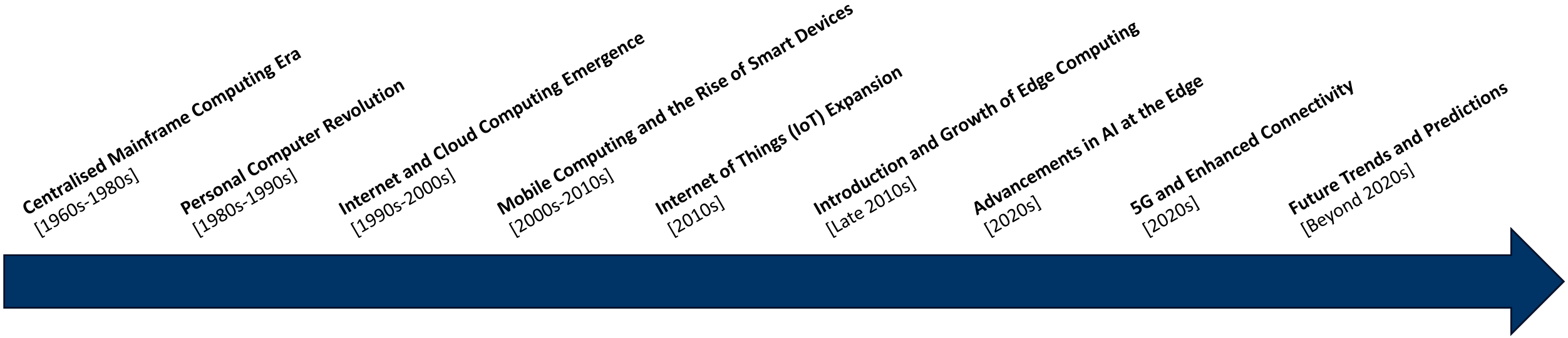
Increased Demand for Low Latency

Data Privacy Regulations

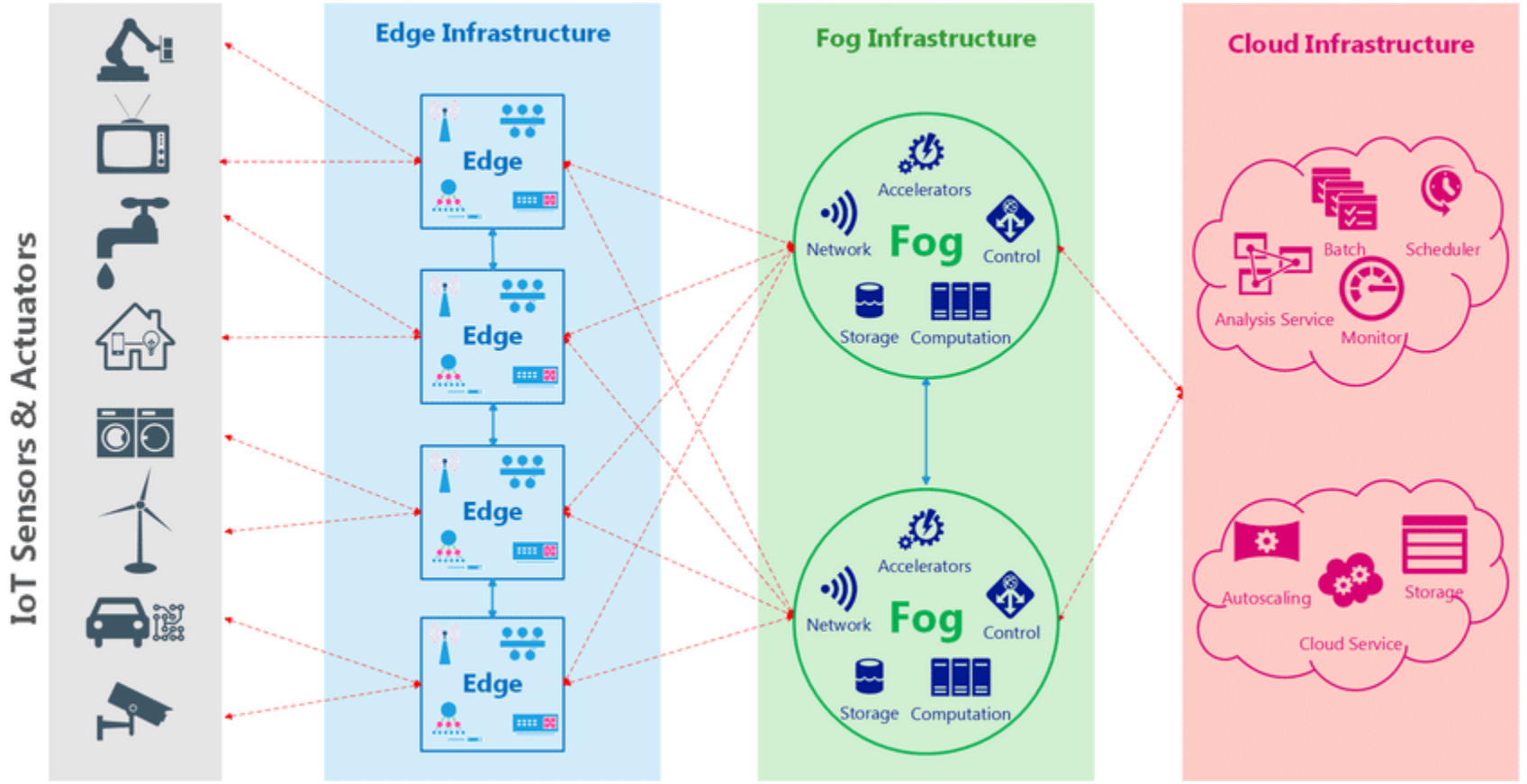
What is Edge Computing?

- Definition: "Edge computing is a distributed computing paradigm that brings computation and data storage closer to the location where it is needed."
- Key Points:
  - Reduces latency
  - Improves response times
  - Saves bandwidth

Evolution of Edge Computing



Edge-Fog-Cloud Architecture



Ref: Cao, Hung & Wachowicz, M. & Renso, Chiara & Carlini, Emanuele. (2019). Analytics Everywhere: Generating Insights From the Internet of Things. IEEE Access. 7. 71749-71769. 10.1109/ACCESS.2019.2919514.

What is Edge Computing

- **General Definition**: Edge computing is a distributed computing paradigm that brings computation and data storage **closer** to data sources such as sensors, devices, and end-users. This proximity **reduces latency** and **bandwidth usage**, **enabling faster processing** and real-time responsiveness.

- **National Institute of Standards and Technology (NIST) Perspective**: Edge computing refers to the **enabling** technologies allowing **computation to be performed** at the **network edge**, on downstream data for cloud services and upstream data for IoT services.

- **Industry Definition**: Edge computing is processing data **near the network's edge**, where the data is generated instead of in a centralised data-processing warehouse. It allows for **efficient data processing** and **analysis**, **reducing the need for data to travel** to distant servers.

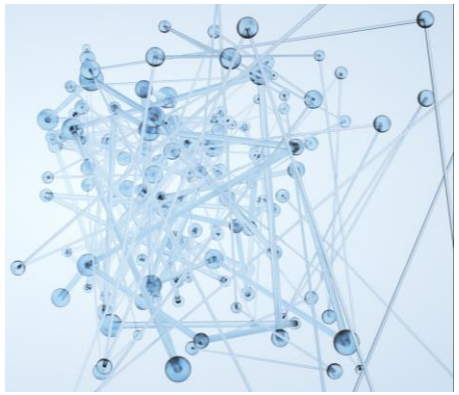
- **Cloud Computing Extension**: Edge computing extends cloud capabilities by distributing processing power **closer to data sources**. It complements cloud computing by handling **time-sensitive data** at the edge while leveraging the cloud for data-intensive and long-term **analytics**.

- **Academic Definition**: Edge computing is a model where data processing and storage are moved away from centralised points to the logical extremes of a network, enabling data **analytics and knowledge generation** to occur at the **source of the data**.

Mapping Between Characteristics and Benefits

CHARACTERISTIC	Leads To	BENEFIT
Local processing of data	**Low Latency** (Enables real-time responses)	Reduced latency for time-sensitive applications
	Enhanced Privacy Measures (Data doesn't leave the device)	Improved data privacy and security
	Autonomous decisions (Immediate, context-aware actions)	Increased responsiveness and relevance
Data filtering at the source	**Reduced Bandwidth Usage** (Less data transmitted over networks)	Bandwidth optimization, cost savings
Resilience to Network Issues	Continues functioning offline	Improved reliability
Resource Constraints	Requires optimized solutions	Energy efficiency, cost efficiency
Heterogeneous Environments	Devices work together seamlessly	Scalability and flexibility

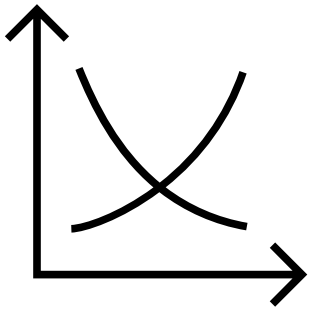
Challenges – Post Deployment



COMPLEXITY

ENVIRONMENT

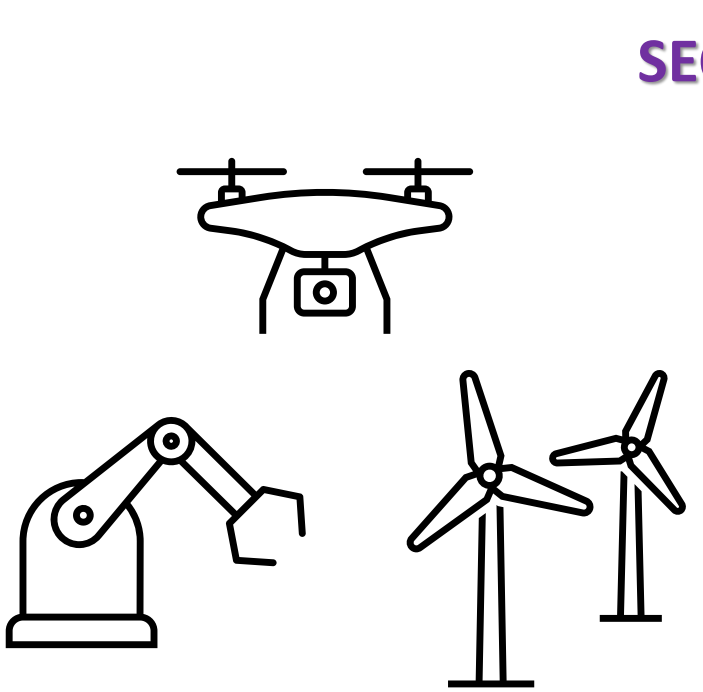
REMOTE LOCATION



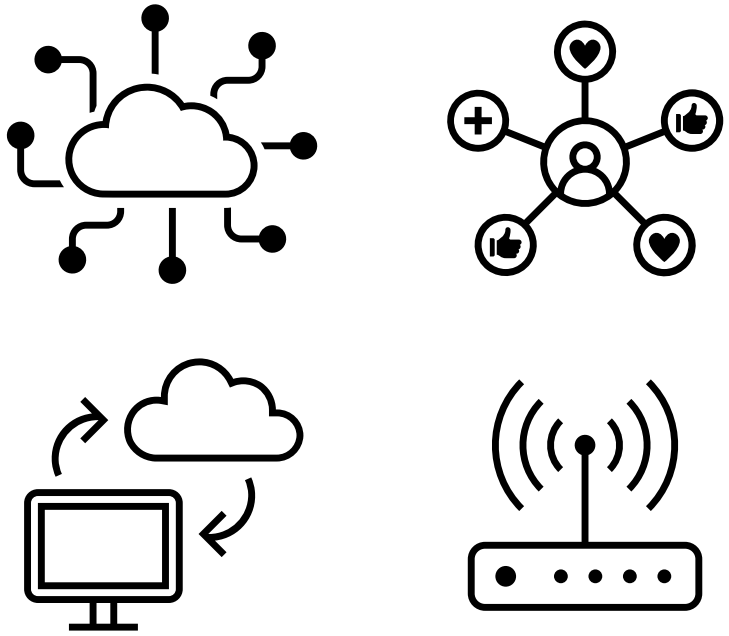
PRICE
vs
REDUNDANCY

Challenges – Operational

SECURITY



SENSOR



NETWORK

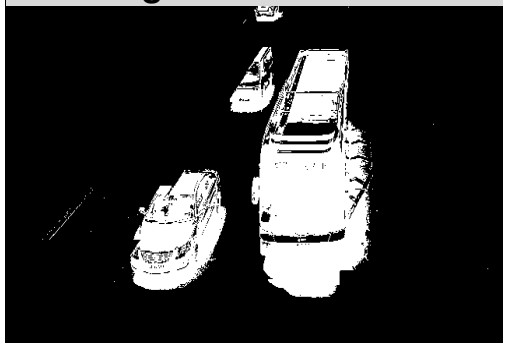
Challenges – Manageability

Holistic tool monitor entire Edge Ecosystem

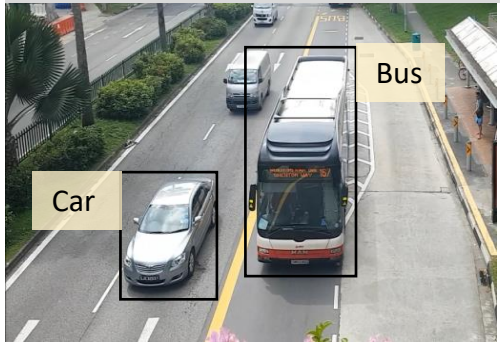
Start Debugging from where?

An Application: Illegal Parking Detection

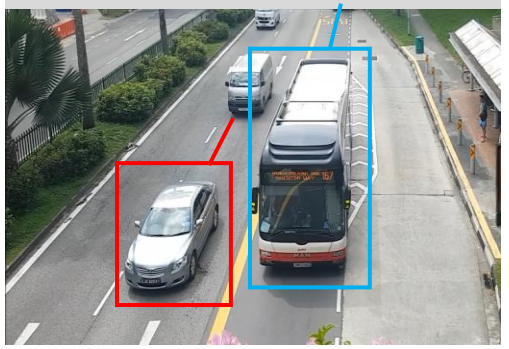
Pre-processing
Background Modeling & Foreground Detection



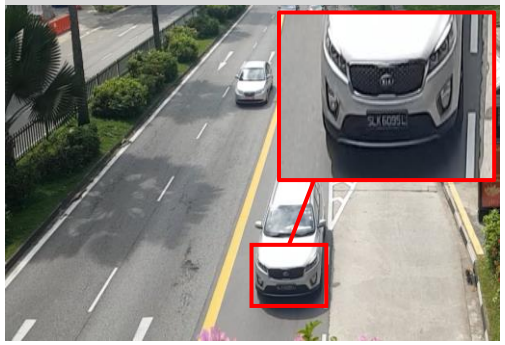
Classification



Vehicle Localization & Tracking

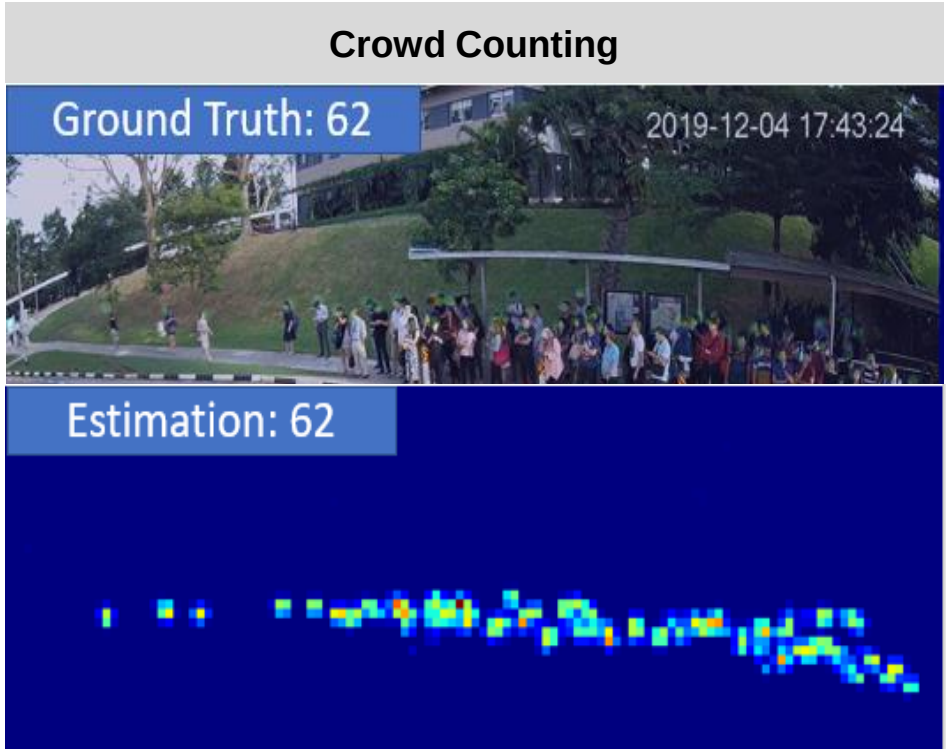


Evidence Collection



Key Characteristics

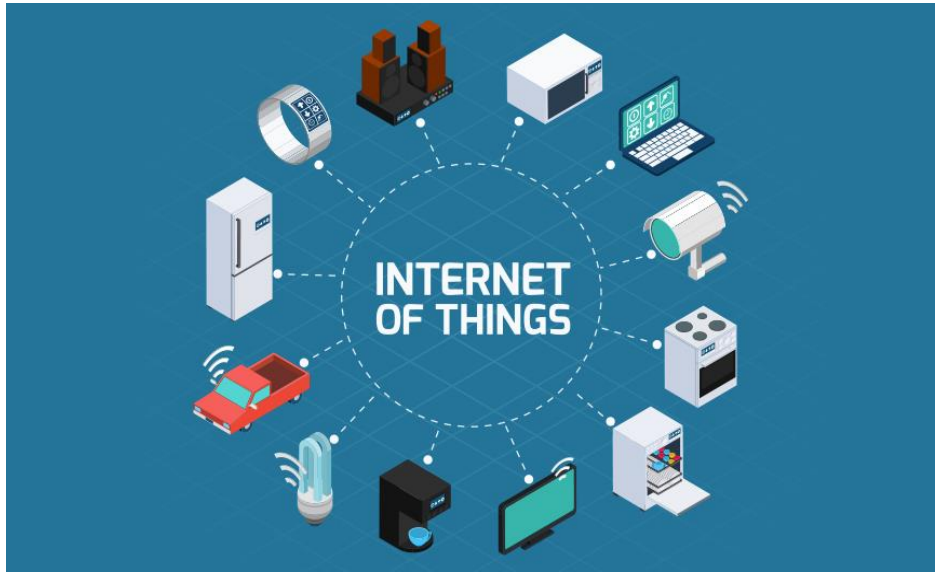
- Processing data near its source.
- Minimizing the distance data travels.
- Reducing latency and improving response times.
- Operating with intermittent connectivity.
- Enhancing privacy and security by localizing data.
- Significantly reduced data storage requirements



AIoT: Fusing AI with IoT – Relies on Edge Computing

SINGAPORE INSTITUTE OF TECHNOLOGY

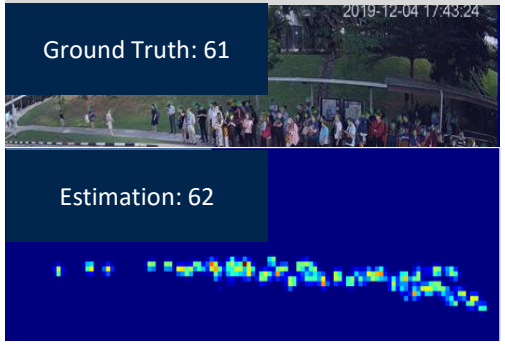
Traditional IoTs



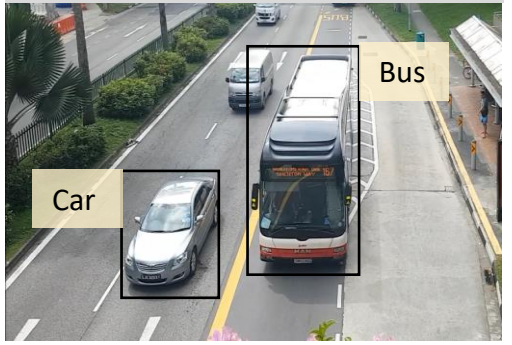
Source: https://www.bankinfosecurity.com/gao-assesses-iot-cybersecurity-other-risks-a-9926

AIoTs

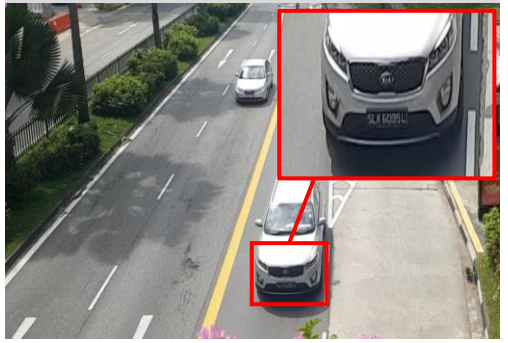
Crowd Counting



Object Localization & Classification



License Pate Recognition



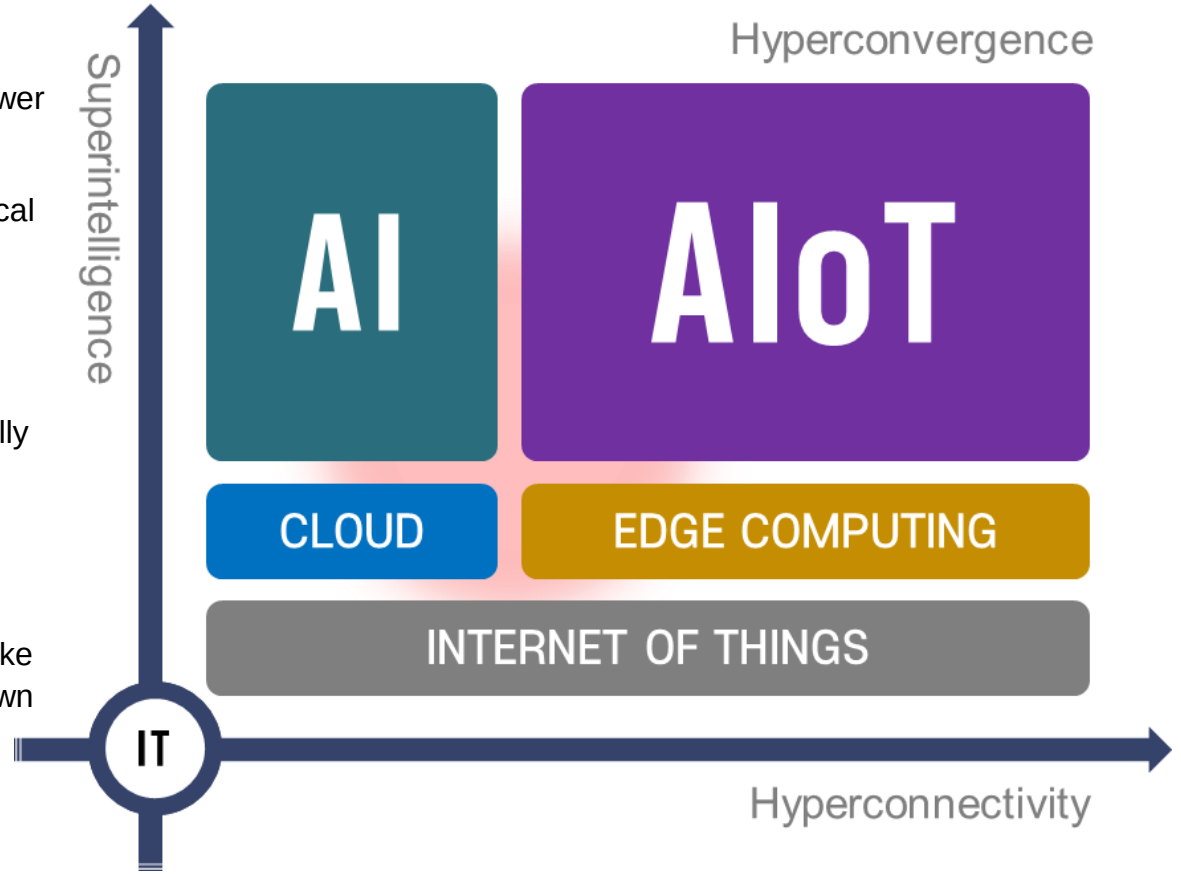
AIoT

- Combines AI (Artificial Intelligence) with IoT.
- Refers to smart IoT devices that use AI technology, such as machine learning and data analytics, to analyse data and make decisions locally at the device or on nearby edge computing infrastructure.
- AIoT can be considered a subset of edge computing where the edge devices are equipped with AI capabilities.
- Key Aspects:
  - IoT devices with embedded AI.
  - Real-time data analysis and decision-making at the device level.
  - Self-learning and adaptive systems.
  - Predictive maintenance and intelligent automation.

- Only the Inference task typically done on the IoT
- Training is still performed offline on desktop class GPUs is typically done

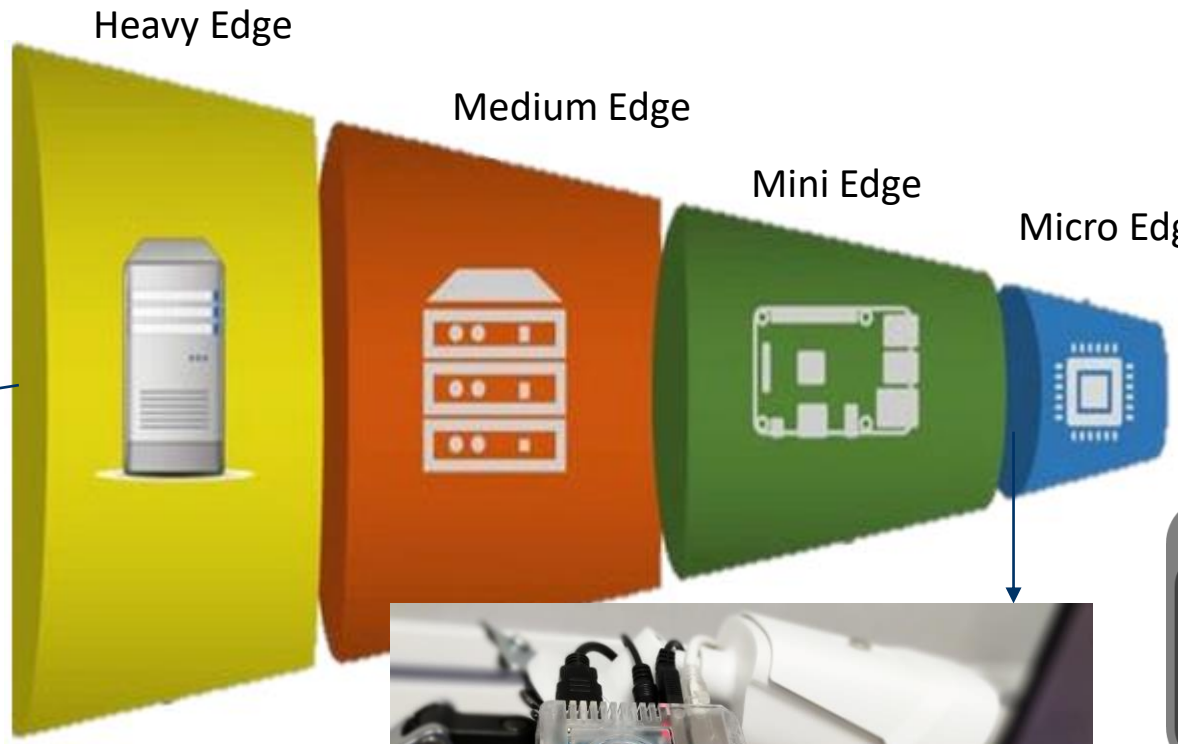
How they all converge

- **AI Integration**: AI enables IoT devices to process data and make intelligent decisions locally, rather than just sending data to a central location.
- **Edge Computing Enabling AI**: Edge computing provides the computational power needed for AI near the data source, reducing latency for real-time processing.
- **Data Processing and Privacy**: Combining AI with edge computing allows for local data processing, enhancing data privacy and reducing breach risks.
- **Efficient Use of Bandwidth**: AIoT devices process and condense data locally, sending only necessary insights to the cloud, saving network bandwidth.
- **Scalability**: Edge computing allows for scalability in IoT by processing data locally rather than overloading cloud servers.
- **Enhanced Functionality**: AIoT devices improve over time, learning from interactions and user preferences to adjust functionality proactively.
- **Real-time Analytics and Decision-making**: AIoT devices analyze data and make real-time decisions at the edge, crucial for immediate responses like shutting down faulty factory equipment.
- **Operational Efficiency**: AIoT predicts maintenance, optimizes resources, and enhances operational efficiency, reacting autonomously to changes.
- **Smarter Infrastructure**: AIoT transforms devices into smart nodes capable of complex processing and autonomous decision-making for smarter environments.

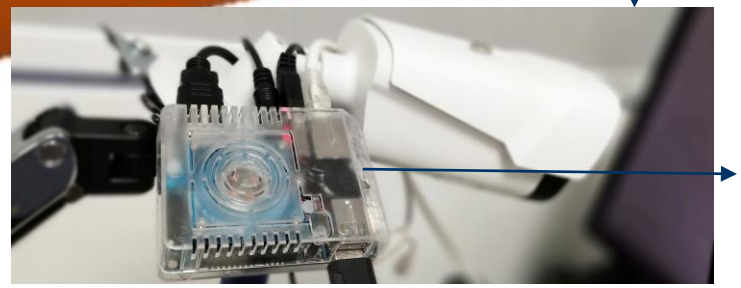


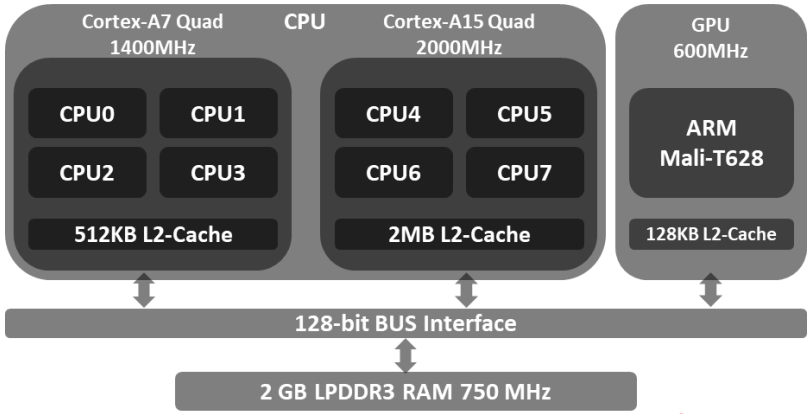
Core Components of Edge Computing

Edge Computing Platforms: Good old Embedded Systems



Credit:
Janakiram.com

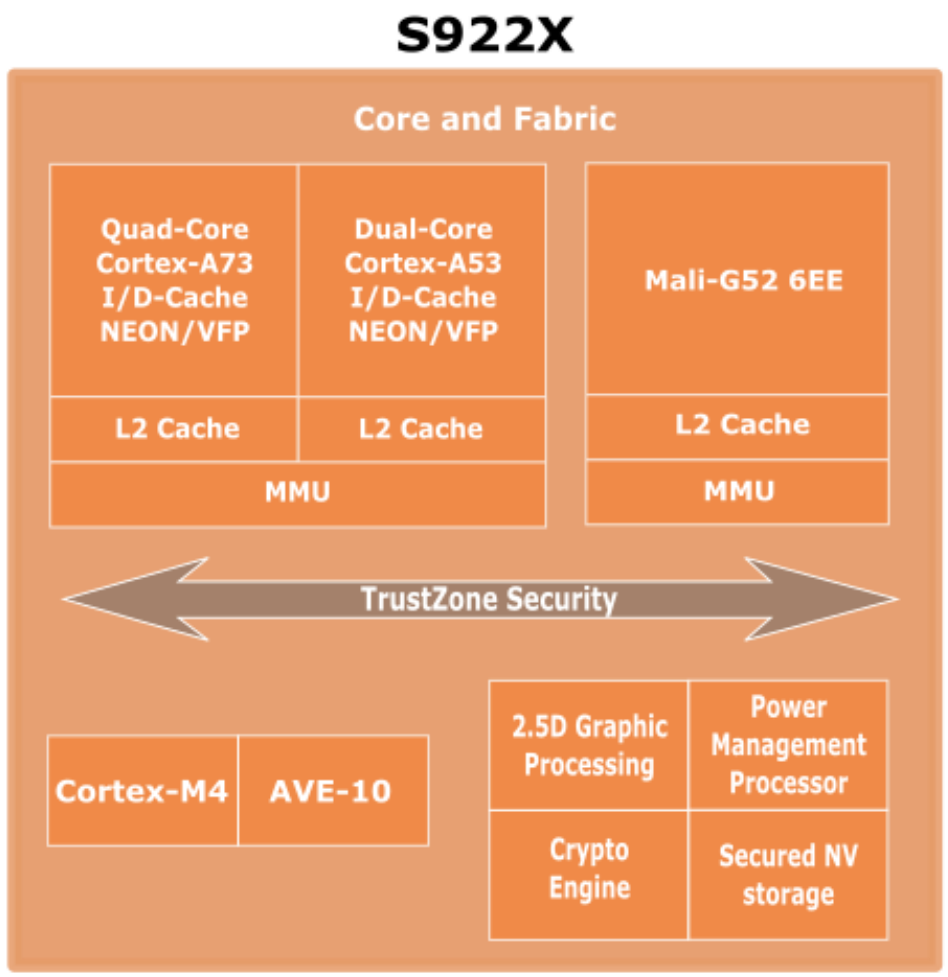




Heterogeneous Edge Computing Platforms

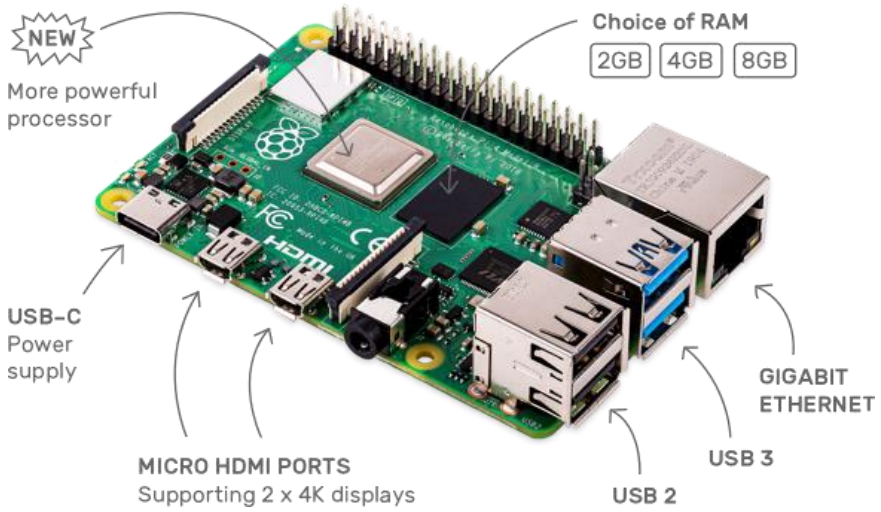
So how do we map our applications effectively on such heterogeneous platforms while considering power and performance constraints?

Edge Devices – CPU centric



6-core CPU from Odroid N2+

Raspberry Pi 4



Specifications

- Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz
- 2GB, 4GB or 8GB LPDDR4-3200 SDRAM (depending on model)
- 2.4 GHz and 5.0 GHz IEEE 802.11ac wireless, Bluetooth 5.0, BLE
- Gigabit Ethernet
- 2 USB 3.0 ports; 2 USB 2.0 ports.
- Raspberry Pi standard 40 pin GPIO header (fully backwards compatible with previous boards)

Edge Devices – GPU centric

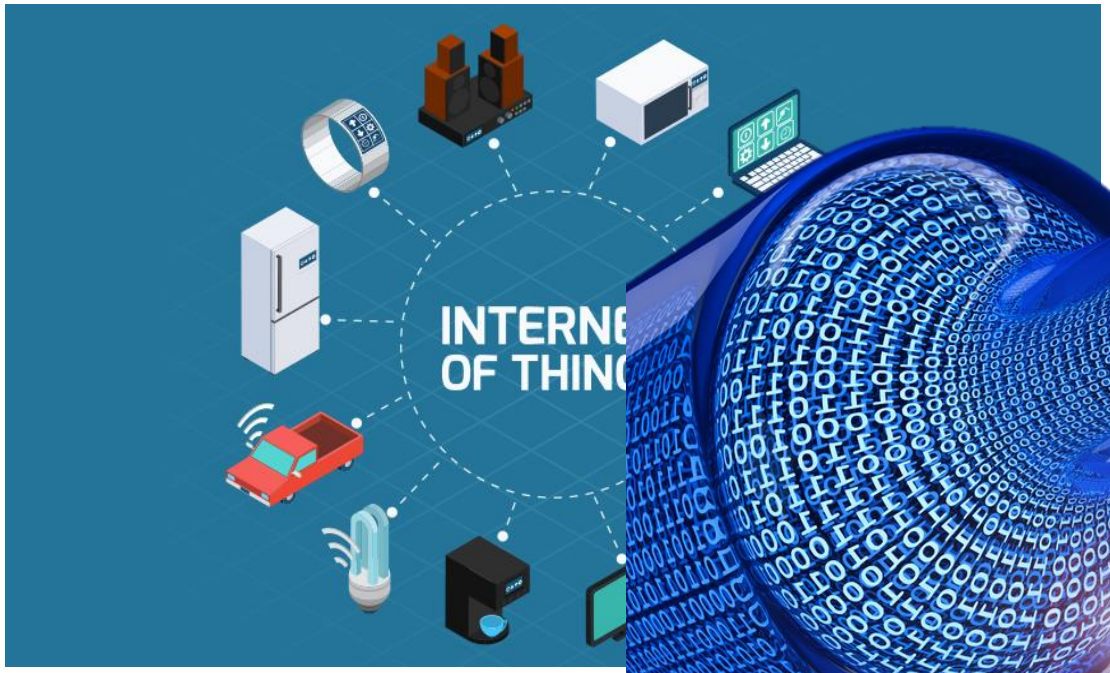
SINGAPORE INSTITUTE OF TECHNOLOGY

Compare NVIDIA Jetson Module Specifications

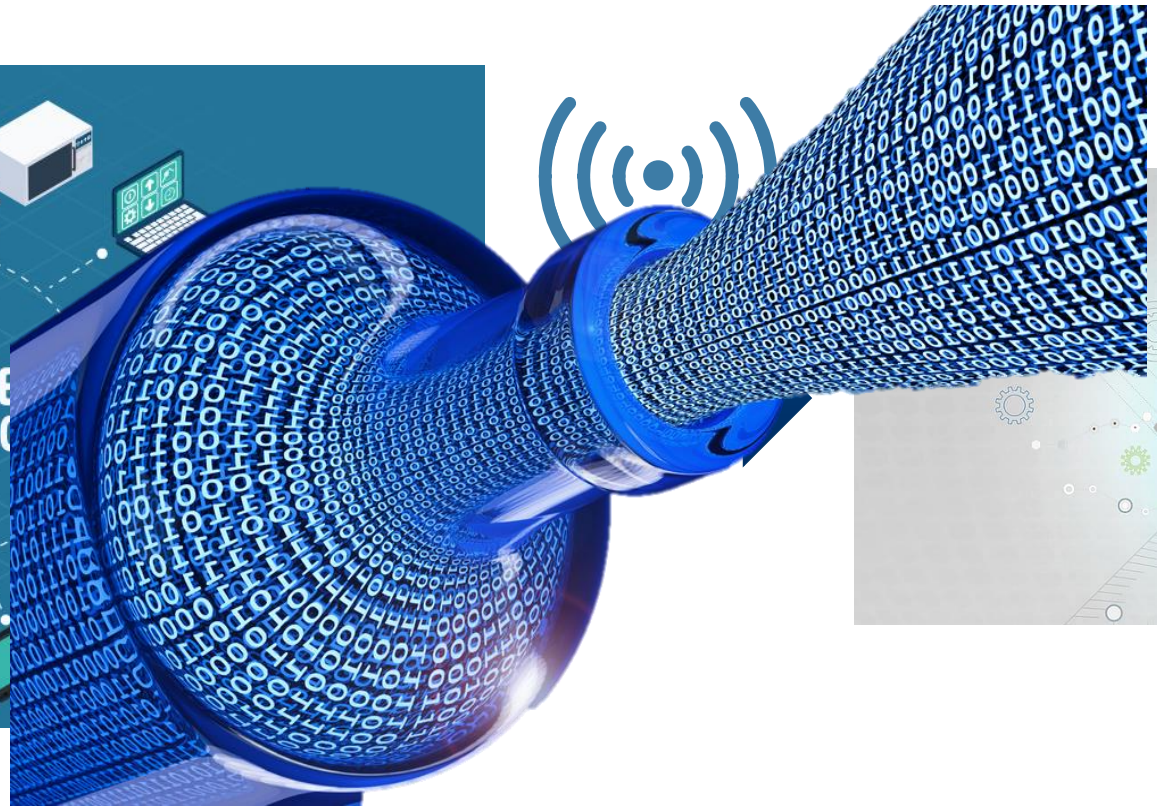
	Jetson Nano	Jetson TX2 Series				Jetson Xavier NX	Jetson AGX Xavier
		TX2 NX	TX2 4GB	TX2	TX2i		
AI Performance	472 GFLOPS	1.33 TFLOPS			1.26 TFLOPS	21 TOPS	32 TOPS
GPU	128-core NVIDIA Maxwell™ GPU	256-core NVIDIA Pascal™ GPU				384-core NVIDIA Volta™ GPU with 48 Tensor Cores	512-core NVIDIA Volta™ GPU with 64 Tensor Cores
CPU	Quad-core ARM® Cortex®-A57 MPCore processor	Dual-core Denver 2 64-bit CPU and quad-core Arm® Cortex®-A57 MPCore processor				6-core NVIDIA Carmel ARM®v8.2 64-bit CPU 6MB L2 + 4MB L3	8-core NVIDIA Carmel Arm®v8.2 64-bit CPU 8MB L2 + 4MB L3
Memory	4 GB 64-bit LPDDR4 25.6GB/s	4 GB 128-bit LPDDR4 51.2GB/s		8 GB 128-bit LPDDR4 59.7GB/s	8 GB 128-bit LPDDR 4 (ECC Support) 51.2GB/s	8 GB 128-bit LPDDR4x 51.2GB/s	32 GB 256-bit LPDDR4x 136.5GB/s
Storage	16 GB eMMC 5.1	16 GB eMMC 5.1		32 GB eMMC 5.1	32 GB eMMC 5.1	16 GB eMMC 5.1	32GB eMMC 5.1
Power	5W | 10W	7.5W | 15W			10W | 20W	10W | 15W	10W | 15W | 30W

Networking in Edge Computing

- IDC estimates that there will be **41.6 billion connected IoT** devices generating a whopping **79.4 zettabytes of data** in 2025[1].



Source: https://www.bankinfosecurity.com/gao-assesses-iot-cybersecurity-other-risks-a-9926

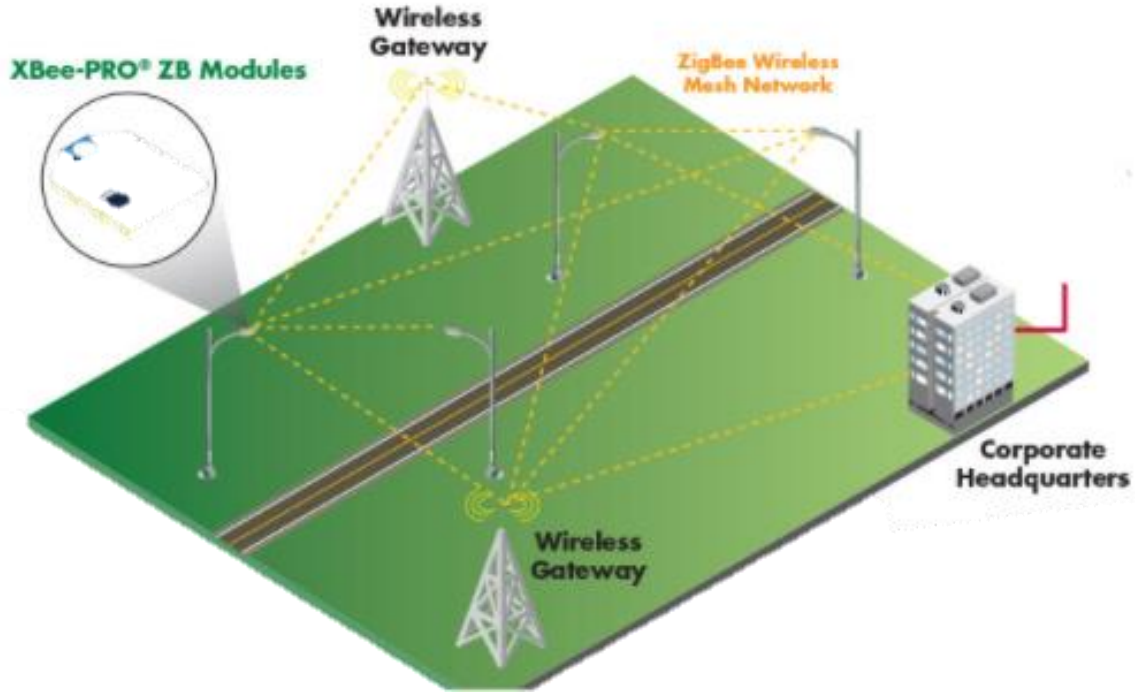


Source: https://www.hpcwire.com/

[1] Worldwide Global DataSphere IoT device and data forecast (2019-2023), IDC

From Edge to Cloud: Networking Options

- Local edges can be connected to a gateway device through a localized interfacing protocol such as ZigBee, Z-wave and Wi-Fi.
- Connectivity to Cloud could be done using LTE, upcoming 5G or the myriad of other long range protocols we studied during this course, depending on the need for bandwidth, amount of data to be sent, cost of networking, etc.



Software and Technologies: Enabling Intelligence at the Edge

SINGAPORE INSTITUTE OF TECHNOLOGY

- Fog Computing
- Containerisation - Docker
- Deployment, scaling, and operations of containerized applications - Kubernetes
- Fleet Management - Balena

Introduction to Analytics in Edge Computing

Data Processing at the Edge

- Local Data Processing:
  - **Definition**: Processing data where it is generated – at the edge of the network, close to IoT devices and sensors.
  - **Benefits**: Reduces latency, minimizes bandwidth use, improves response times, enhances privacy and security.

- Edge vs. Traditional Analytics:
  - **Traditional Analytics**: Often involves sending data to centralized servers or cloud data centers for processing.
  - **Edge Analytics**: Data is analyzed on the edge device itself or on a nearby edge server.
  - **Comparison**: Highlight the reduced latency and real-time decision-making capability of edge analytics versus the more resource-intensive and slower traditional approach.

EDGE AI ECOSYSTEM

SINGAPORE INSTITUTE OF TECHNOLOGY

DNN network architecture

- Classification (VGG, Mobilenet, Squeezenet)
- Object detection (faster-RCNN, YOLO, SSD)

DNN framework (training,inference)

Caffe, Caffe2, Pytorch (Facebook), Tensorflow (Keras)(Google), MXNet, Darknet

Inference-only DNN framework

TensorflowLite, NCNN, Pytorch Mobile

(Weight Compression, Filter Pruning, Quantisation)

CV and ML Hardware Libraries

- ARM NN and ARM Compute Library

CPU

- ARM big.LITTLE

GPU

- Nvidia Jetson
- ARM Mali

SoC Families

Hisilicon Kirin, Samsung Exynos, Qualcomm Snapdragon, Nvidia Jetson, Apple Bionic

FPGA/Accelerator

Google edge TPU (coral), Intel Nirvana NNP, Apple Neural Engine, Huawei NPU (in Kirin SoC)

Hardware Programming Languages

Multi-core CPU

- OpenMP, NEON (extension)

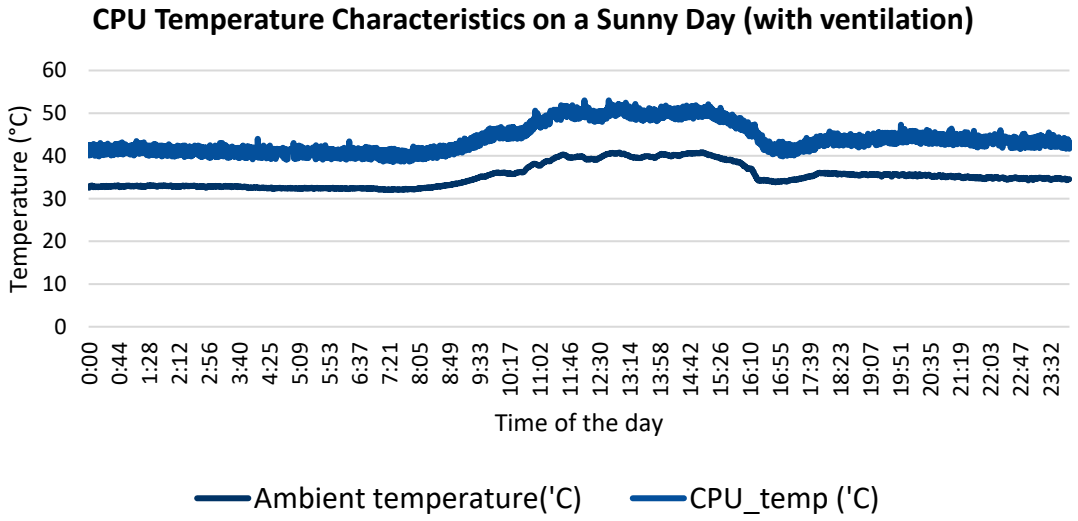
GPU

- CUDA, OpenCL, cuDNN

Edge Analytics in Action

- Waterproofing requirements
- Thermal Challenges
- Waterproofing and thermal are conflicting requirements
- Temperature inside the box can rise over 65°C during daytime, without any ventilation
- And CPU Temperature is even higher than that. Can easily rise to 85°C during daytime

Traffic Junction



Challenges and Future Trends

Challenges in Edge Computing

- **Security**: Edge computing presents unique security challenges due to the distribution of data processing across numerous devices.
- **Computational Limitations**: There are inherent limitations in the computational capabilities of edge devices compared to centralised data centres.
- **Data Management**: Managing and coordinating data across various edge computing devices poses significant challenges.

Smart Manufacturing Examples: Airbus – Factory of the Future

- MiRA (Mixed Reality Application) tablet
  - Cross between a sensor pack and a tablet

- Internet Connected Smart Tools
  - Auto-adjust to different actions
  - Log information
  - Reduces assembly time

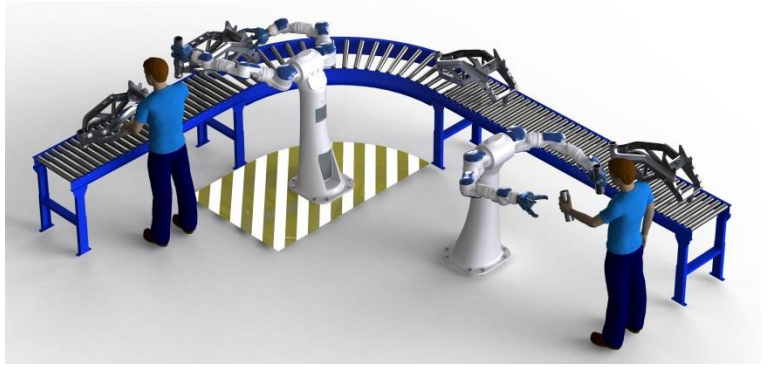
- Augmented Reality driven instructional & educational tutorials

Photos courtesy of Airbus Factory of the Future

Smart Manufacturing Examples: Continental AG's SMART Factory

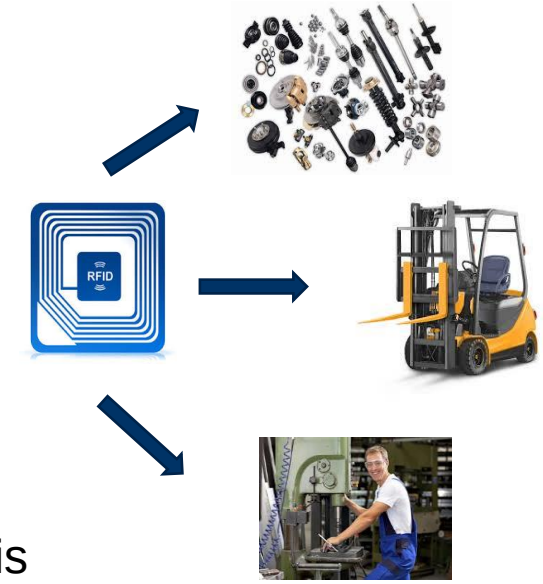
- Active RFID tags and Geo-location are used to move the tire components throughout the factory

- Collaborative robots
  - Robots are "shown" how to do a task once and then they can repeat that action
  - Reduces risks of injuries and reduces the need for additional assisting employees

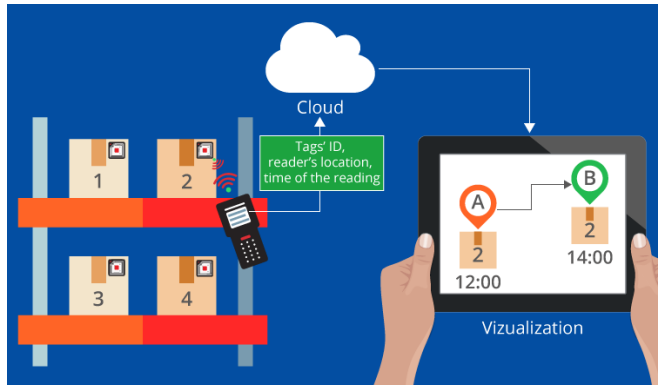


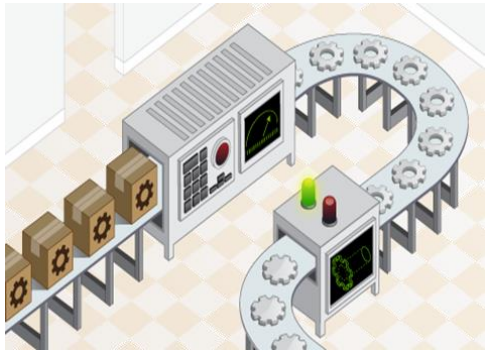
Source: Entrigna

Going forward



- Streamlined Factories
  - Asset tagging, locating, supply chain management
  - Easier to do Just-in-time asset management
- SMART Inventory management
  - Sensors on containers can determine when a product is running low
  - Automatic alerts to proactively re-order the parts or orders can be automatically placed with suppliers
- SMART Quality control
  - RFIDs attached to products can be used to tag defective products
  - Automatic alarms if failure rate crosses a certain threshold for early course correction





Source: Entrigna

Designing for the Edge: Key Considerations

SINGAPORE INSTITUTE OF TECHNOLOGY

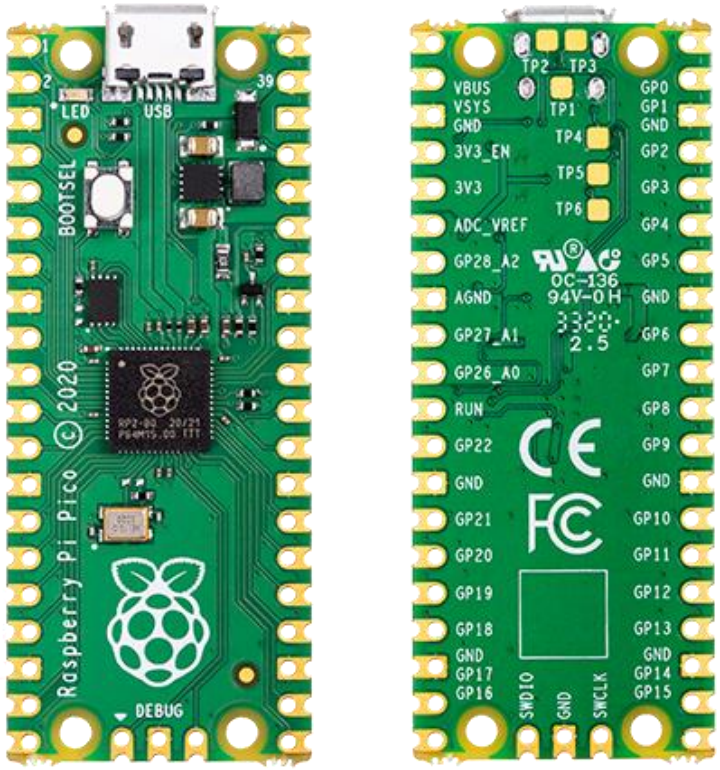
- Memory Constraints
  - Models must fit within the limited RAM and flash storage of microcontrollers.
- Processing Power
  - Algorithms should have low computational complexity to enable real-time inference.
- Energy Consumption
  - Models need to be energy-efficient, especially for battery-powered devices.
- Data Types
  - Use of integer-only arithmetic is common to reduce computational load.
- Model Size
  - Target model sizes are often in the range of tens to hundreds of kilobytes.

Designing for the Edge: Other Considerations

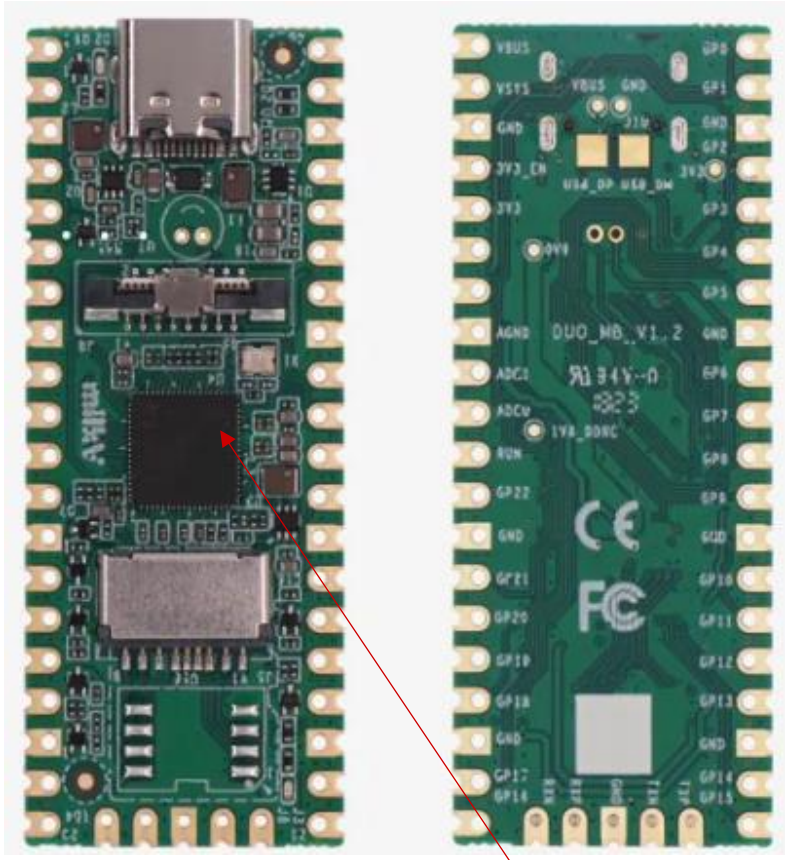
1. Latency and Response Time
2. Scalability - Number of devices and the Geographical
3. Security and Privacy, i.e. privacy concerns
4. Network Connectivity and Bandwidth
5. Data Management and Storage
6. Robustness and Reliability
7. User Interface and Experience
8. Regulatory Compliance
9. Cost-Effectiveness
10. Customization and Flexibility

Demand for Intelligence at the Edge

Raspberry Pi Pico



Milk V Duo



The inbuilt Tensor Processing Unit (TPU) delivers 1.0 TOPS (Tera Operations Per Second) of computing power for 8-bit integer operations, significantly improving performance for efficient machine learning inference tasks.

Summary

- What is it and its characteristics
  - AIoT: Artificial Intelligence Tasks on IoT Devices
- Core Components of Edge Computing
  - Several Edge Computing Platforms; CPU vs GPU
  - Networking; MQTT, CoAP, etc
  - Architecture; Cloud, Fog, etc
- Analytics in Edge Computing
- Challenges