White Paper

AIoT - Memory and Storage Solutions for AI at the Edge

Executive Summary

Artificial Intelligence (AI) and the Internet of Things (IoT) are in a process of merging into what we define as AIoT. With edge computing, computational power is moving to the edge where IoT devices are gathering data. AI is the next logical step in efficient data handling and lowering latency while opening up for innovative solutions at the edge.

Conditions on the edge where this AI computation takes place are varied and any device has to account for this. This paper explains these trends in light of the need for optimized storage and memory solutions for these AI edge applications.

Introduction

We are moving into a new era of technological innovation. The concept of the Internet of Things (IoT) has already been around for a long time, especially so in the light of our rapid technological development. IoT incorporates the spirit of physical and digital convergence; data is gathered from an increasing number of devices for then to be aggregated into what is commonly known as Big Data. The number of these devices continues to grow and is estimated to reach a staggering 50 billion by 2020.

The data gathered by these devices encounter a problem when attempting to transmit to a centralized location such as the cloud, namely latency. Even though connection speeds are steadily increasing it fails to keep pace with the exponentially growing amount of data. Unless handled, this means that latency will increase and overall system performance will suffer.

This is one of the areas where AI can make significant contributions. Furthermore, it also opens up new technological innovations such as streamlining city traffic to public security and enhanced financial services.

More fundamentally, AIoT requires components that can handle the challenging and diverse conditions found at the edge. These locations can be anything from onboard vehicles and airplanes to factories or oil installations in the desert. This requires a flexible and adaptive approach to component manufacturing. AI also promises to reduce the human factor when it comes to decision making. This puts greater pressure on system integrators to ensure quality control as an accident involving AI, where the human factor is removed, will not necessarily have a clear and obvious culprit.

Background

Let us first define the concepts of IoT, AI, and edge computing:

IoT

The internet of things is a phrase that refers to the trend of "things" being interconnected through a network (usually the internet). The "things," in this case, do not necessarily refer to separate electronic devices; they can also refer to things like wearable electronics or even people that have a medical device on or implanted in them. Basically, it is every individual device that can transfer data within a network in some fashion.

AI

The AI we are referring to fits within the concept of "Narrow AI." This is a program or system that is able to perform a set of specific tasks without any direct human input on how to do so. This differs significantly from "General AI," which is the AI we are used to seeing in movies and series which has human-like autonomous capabilities. A current example of narrow AI is text, picture, and speech recognition that we can create through neural networks and machine learning. Such an AI has gone through thousands, if not millions, of different data iterations and taught itself how to correctly identify the image or object at hand.

But no matter how sophisticated its predictions become, it is still limited to this narrow function it has been trained for. If anything falls outside of this scope, the AI is rendered all but useless. An AI trained to identify written numbers can learn its task and will easily supersede human capabilities, but it will be completely useless when given a task such as identifying letters.

Edge Computing

The original idea of IoT had data sent to a central location, or the cloud, to undergo processing and analysis. However, as the number of devices has increased exponentially, many applications have reached a roadblock where this large amount of data transmitted back and forth causes severe latency issues.

Edge computing tackles this problem by handling more data at the edge. This way the device can determine by itself what needs to be sent to the cloud and what can be filtered out. The concept simply means moving computational power out to the "edge," where the internet connects to various devices, i.e. the location data is actually gathered.

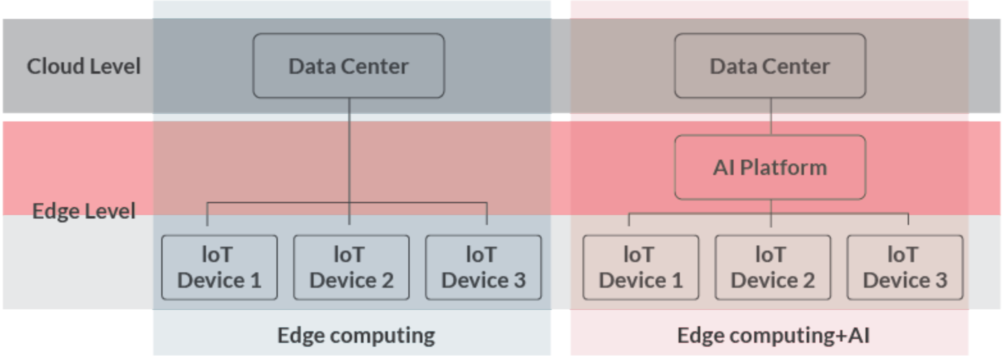

Figure 1: Edge computing adds an additional layer to the edge level with the aim to reduce latency

Challenges

Limits of IoT

IoT in its pure form only gathers data with little or any computation. This means that the data is sent in bulk to cloud to be analyzed. However, all data is not equally valuable. Take for example security footage, the interesting parts have people or objects moving, while still shots of an unchanging background are less interesting. In this case, sending all the data to a cloud for analysis would waste large amounts of bandwidth that could have been used for other applications.

Computational Power and Harsh Environments

AI at the edge can potentially demand a lot of computational power to ensure performance is adequate. However, standard storage and memory components might deliver the needed performance but are ill-equipped to handle the rough conditions at their specific location. E.g., road-side traffic monitoring will experience temperature cycles from day to night and summer to winter, in-vehicle systems have to contend with shock and vibration, industrial settings have increased levels of pollution etc.

Solutions

The AI Platform

When talking about AIoT, we usually refer to an AI platform located at the edge. This normally takes the form of a small IPC with an inbuilt industrial-grade CPU. For real-time data analysis, this CPU needs adequate support in form of flash memory and DRAM.

Industrial-grade Memory and storage

Industrial-grade storage and memory components are essential to solving the issues of implementing AI at the edge. The main issues to solve are exploring and identifying the risks present at each location of data gathering. The components can then be customized to fit the specific requirements of the application. Let us look at some examples of how this would work out in real-life scenarios:

City Traffic Surveillance

Our cities are growing in three dimensions by spreading outward and upward (by buildings growing in height). Roads, however, are still mostly confined to two dimensions, which leads to increased traffic congestion as cities grow larger.

Monitoring and altering traffic flow based on real-time data can significantly increase efficiency and cut down congestion. This can be done with surveillance installations strategically placed throughout the city.

The first-step analysis is handled by local AI platforms at the edge. This includes vehicle recognition and traffic flow assessment. Each installation can thus determine by itself how to handle the data based on the analysis; i.e., is the number of vehicles increasing and is there a risk of congestion? Any essential data can then be sent to a centralized platform (the cloud), where measures such as redirecting traffic, altering speed limits, and adjusting traffic lights can be taken based on the data.

Fleet management and AI

AI can significantly optimize fleet management operations. Monitoring a large fleet of vehicles can be hard but there are many ways to streamline operations: reducing fuel costs, vehicle maintenance, mitigating unsafe driver behavior etc.

The current positioning systems are mostly reliant on GPS, which fails to handle certain problems. For example, entering a tunnel renders the GPS all but useless and the system will have no idea where the vehicle is located. This also happens within cities when driving inside buildings or other areas with poor satellite coverage. It is also difficult for the system to determine the vehicle's elevation.

However, there are other sources of data other that can give us a pointer on vehicle position: Firstly, a vehicle's speed and turning rate can be constantly monitored and logged. An onboard AI platform can then calculate the vehicle position is at any point in time by having these parameters compensate for incomplete GPS data. This technology is called automotive dead reckoning, or DR. Lastly, data can be transmitted through wireless networks back to the operator.

Autonomous Delivery Robots

When we remove the human factor from vehicles, the main problem we run into is the ever-changing traffic picture that is fraught with unexpected factors. Because of this, an autonomous vehicle has to be able to make split-second decisions with any sudden change happening in its path. Where we rely on our senses, the robot has a multitude of sensors that gather all kinds of data that has to be processed into a coherent image of the overall situation at any moment in time. Relying on the cloud, in this case, is hopeless as the latency will surely mean that by the time the data is ready and a decision can be made it is already too late.

The onboard AI platform that handles these complex calculations is reliant on components that work under whatever weather and physical conditions are present without any drop in performance. To avoid accidents involving autonomous vehicles it is prudent that the equipment is performing with minimal chance of failure and with sufficient backup.

Conclusion

AI is here to stay and as its role in IoT grows we have to look for smart solutions that ease this transition. Furthermore, AI is poised to supplant the human operator in many scenarios which further stresses the need for robust systems that can handle any relevant environmental challenge.

Powering AI platforms with industrial-grade memory and storage solutions is the way to ensure that the hardware is up for the task and is one of the key components in building the IoT of the future.

Innodisk Corporation

5F., NO. 237, Sec. 1, Datong Rd., Xizhi Dist., New Tapei City, 221, Taiwan
Tel : +886-2-7703-3000
Fax : +886-2-7703-3555
E-Mail : sales@innodisk.com
Website : www.innodisk.com

innodisk