White Paper

Memory and Storage Solutions for AI at the Edge

Executive Summary

Artificial Intelligence (AI) and the Internet of Things (IoT) are in the 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 improving the efficiency of data processing and lowering latency while also allowing for innovative solutions at the edge.

Additionally, to ensure operational stability and device reliability, connected devices at the edge responsible for such AI computations also need to account for the environmental conditions in their vicinity.

This white paper outlines the AI and IoT trends and their merging into AIoT as well as the necessity of optimized storage and memory solutions to enable the AIoT edge applications of the future.

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 considering the rapid pace of technological development.

Today, IoT represents physical and digital convergence; an increasing number of devices gather data before aggregating it into what is commonly known as *big data*. The number of these connected devices continues to grow and is estimated to reach a staggering 50 billion by 2020.

However, these devices encounter a problem when attempting to transmit the data they have gathered to a centralized location such as the cloud, namely latency. Even though the quality of internet connections is steadily increasing, it fails to keep pace with the exponentially growing amount of data that our devices generate. Unless handled, this means that latency will increase, and overall system performance will suffer.

Latency is one of the areas where AI can make significant contributions.

Furthermore, AI also invites new technological innovations such as streamlining city traffic, improving public security, and enhancing financial services.

On a more fundamental level, AIoT requires components that can handle the challenging and diverse conditions found at the edge. These conditions can be as varied as onboard vehicles and airplanes to factories or oil installations in the desert. Such diversity requires a flexible and adaptive approach to manufacturing components.

Moreover, AI promises to reduce the human factor when it comes to decision-making. This development puts greater pressure on system integrators to ensure quality control as an accident involving AI, where the human element is removed, will not necessarily have a clear and obvious culprit.

Background

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

The Internet of Things

The internet of things is a phrase that refers to the trend of connecting things through a network (usually the internet). The “things,” in this case, does not necessarily refer to separate electronic devices; it can also refer to things like wearable electronics or even people that have medical devices on or implanted in them. In essence, it refers to every individual device that can transfer data within a network in some capacity.

Artificial Intelligence

The AI we are referring to here fits within the concept of "narrow AI." Narrow AI is a program or system that can perform a set of specific tasks without any direct human input on how to do so. This type of AI differs significantly from "general AI," which is the AI we are used to seeing in movies and TV shows that has humanlike autonomous capabilities.

In contrast, an example of narrow AI is text, picture, and speech recognition that we can create through neural networks and machine learning. Such AI has gone through thousands, if not millions, of different data iterations and taught itself how to identify the image or object at hand correctly. But, no matter how sophisticated its predictions become, this AI is still limited to this narrow function that it was trained for. If anything falls outside of this scope, the AI is rendered all but useless. In other words, AI trained to identify written numbers can learn its task and will easily supersede human capabilities. Still, it will be entirely at a loss when given a task such as recognizing letters.

Edge Computing

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

Edge computing tackles this problem by handling more data at the edge, i.e., where the data is gathered in the first place. This way, the device can determine by itself what it needs to send to the cloud and what it should filter out. While it requires moving more computational power to the edge, it effectively tackles the latency challenge in IoT applications.

Challenges

Limits of IoT

IoT, in its pure form, gathers data with little or without any computation. What this means is that the data it collects is sent in bulk to the 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 practically still shots of an unchanging background are less interesting. In this case, sending all this data to the cloud for analysis would waste substantial amounts of bandwidth that other applications otherwise could have used.

Computational Power and Harsh Environments

AI at the edge can potentially demand a lot of computational power to ensure that performance is adequate. However, while standard storage and memory components might deliver the needed performance, they are often ill-equipped to handle the rough conditions where they are expected to operate. Components used in road-side traffic monitoring, for example, 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, and so on.

Solutions

The AI Platform

When talking about AIoT, we usually refer to an AI platform located at the edge. In practice, this normally is an industrial PC (IPC) with a built-in industrial-grade CPU. To allow real-time data analysis, this CPU needs adequate support from other components, e.g., flash memory and DRAM.

Industrial-grade Memory and Storage

Industrial-grade storage and memory components are essential to solving the challenges in implementing AI at the edge. The main issues to solve are exploring and identifying the risks present at each location where data gathering takes place. These components can then be customized to fit the requirements of their specific 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 upward and outward with taller city centers and ever-growing suburbs. Roads, however, are still mostly confined to two dimensions, which leads to increased traffic congestion as our cities keep growing.

Monitoring and altering traffic flow based on real-time data can significantly increase efficiency and reduce congestion. This functionality can be accomplished by strategically placing surveillance installations throughout the city.

Local AI platforms handle the first-step analysis at the edge. This analysis includes vehicle recognition and traffic flow assessment. Each installation can thus determine, by itself, how to interpret the data based on its analysis; i.e., is the number of vehicles increasing, and is there a risk of congestion? It can then send any essential data to a centralized platform (e.g., the cloud), where measures such as redirecting traffic, altering speed limits, and

managing traffic signals can be taken based on the edge device's data and analysis.

Fleet management and AI

AI can provide significant optimizations for fleet management operations. Monitoring a large fleet of vehicles can be hard, but there are many ways to streamline operations, for example, by reducing fuel costs and vehicle maintenance, mitigating unsafe driver behavior, and so forth.

Modern positioning systems mostly rely on GPS, which fails to handle specific problems. For example, entering a tunnel renders the GPS all but useless, leaving the system without any idea of the vehicle's location. Such issues also occur within cities, for example, when driving inside buildings or in other areas with inadequate satellite coverage. It is also difficult for the system to determine the vehicle's elevation when relying on GPS data.

However, there are other sources of data that help determine the vehicle's position. Firstly, a vehicle's speed and turning rate can be continuously monitored and logged. With that information, an onboard AI system can then calculate the vehicle's position at any point in time by using these parameters to compensate for incomplete GPS data. Finally, this data can be transmitted through wireless networks back to the operator.

Autonomous Delivery Robots

When we remove the human factor from delivery vehicles, the main problem we run into is the ever-changing traffic patterns that are fraught with unknown variables. Because of this, an autonomous vehicle has to be able to make split-second decisions when sudden changes happen in its path. Where humans rely on our senses, autonomous delivery robots have a range of sensors that gather a variety of data that it then processes into a coherent image of the overall situation at any moment in time.

Unfortunately, relying on the cloud is insufficient in this case as the latency causes the reaction time to be too slow for split-second decisions.

While an onboard AI platform can effectively handle these complex calculations, it nevertheless requires components that work under any weather and physical conditions that are present—and without any drop in performance. To avoid accidents involving autonomous vehicles, it is therefore prudent that the equipment is performing with minimal chance of failure and with sufficient security measures in place.

Conclusion

AI is here to stay and, as its role in IoT grows, we have to look for smart solutions that ease this transition. Since AI is poised to supplant human operators in many scenarios with operationally stressful environments, it further underlines the need for robust systems that can handle challenges like vibration, extreme temperatures, moisture, and contaminants.

Therefore, powering AI platforms with industrial-grade memory and storage solutions is the best way to ensure that the hardware is up for the task—making these components key for building the AIoT systems 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