Welcome to AI for Telco Part 6: AI Cloud and Infrastructure

ANI_113d | On-Demand | Automation and Insights | Express
Course Duration: 1 hour

This course explores the growing demand for AI cloud infrastructure, emphasizing scalability, flexibility, and cost efficiency. Participants will compare AI cloud with traditional IT cloud infrastructure. The course provides an overview of AI hardware, including CPUs, GPUs, DPUs, NPUs, and TPUs. A deep dive into CPU architectures covers traditional designs, limitations for AI, and advancements, along with FPGAs and ASICs. The course also examines how GPUs work, their advantages over CPUs for AI, and optimization techniques. Additionally, it covers private cloud solutions for AI training and inference, supported by case studies.

Intended Audience

Those looking for basic introduction to AI cloud and infrastructure

Objectives

After completing this course, the learner will be able to:

- Explain AI cloud demand and benefits
- Compare AI cloud vs. IT cloud
- Identify types of AI hardware
- Analyze CPU architectures and limits
- Evaluate CPU advancements, FPGAs, and ASICs.
- Describe GPU functions and advantages
- Implement private cloud for AI training

Outline

1. Demand for AI Cloud
1.1 Growing need for AI infrastructure
1.2 Scalability and flexibility
1.3 Cost efficiency
1.4 Compare AI cloud vs IT cloud infrastructure

2. Overview of AI Hardware
2.1 Central Processing Units (CPUs)
2.2 Graphics Processing Units (GPUs)
2.3 Data Processing Units (DPUs)
2.4 Neural Processing Units (NPUs)
2.5 Tensor Processing Units (TPUs)

3. Deep Dive into CPU Architectures and Augmenting Hardware
3.1 Traditional CPU Architectures
3.2 Limitations of CPUs for AI
3.3 Advancements in CPU Design
3.4 Field-Programmable Gate Arrays (FPGAs)
3.5 Application-Specific Integrated Circuits (ASICs)

4. Deep Dive into GPUs
4.1 How GPUs Work
4.2 GPUs vs. CPUs for AI
4.3 Optimizing AI with GPUs

5. Private Cloud for AI
5.1 Private Cloud for Training
5.2 Private Cloud for Inference
5.3 Case Studies

6. Summary

© 2024 Award Solutions, Inc.