8. Atlas AI Computing Platform

Cristiano Bacelar de Oliveira

Setembro/2020

Index

1. **Overview of AI Chips**
2. **Hardware Architecture of Ascend Chips**
3. **Software Architecture of Ascend Chips**
4. **Huawei Atlas AI Computing Platform**

8. Atlas AI Computing Platform

8.1 Overview of AI Chips

Cristiano Bacelar de Oliveira

Setembro/2020

Index

Overview of AI Chips

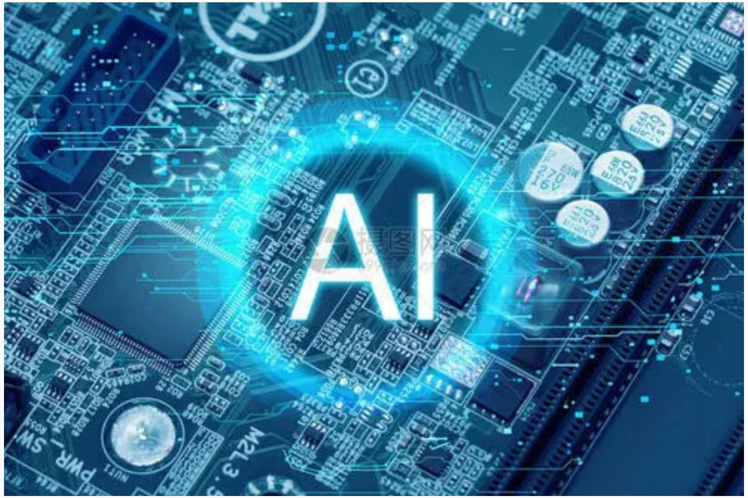
- Classification of AI Chips
- Current Status of AI Chips
- GPU, TPU, and Ascend 310 Design Comparison
- Ascend AI Processors

Disclaimer

The following content is heavily based on HCIA-AI Course material by Huawei Technologies Co., Ltd., authored by Shu Xiaodong. Distribution is not allowed.

Introduction

- Four elements of AI: data, algorithm, scenario, and computing power
- AI chips, also known as AI accelerators, are function modules that process massive computing tasks in AI applications.



Classifications of AI Chips

Classification of AI Chips (1)

- **AI Chips can be divided into four types by technical architecture:**
  - A central processing unit (CPU): a super-large-scale integrated circuit, which is the computing core and control unit of a computer. It can interpret computer instructions and process computer software data.
  - A graphics processing unit (GPU): a display core, visual processor, and display chip. It is a microprocessor that processes images on personal computers, workstations, game consoles, and mobile devices, such as tablet computers and smartphones.
  - An application specific integrated circuit (ASIC): an integrated circuit designed for a specific purpose.
  - A field programmable gate array (FPGA): designed to implement functions of a semi-customized chip. The hardware structure can be flexibly configured and changed in real time based on requirements.

Classification of AI Chips (2)

- **AI chips can be divided into training and inference by business application.**
  - In the training phase, a complex deep neural network model needs to be trained through a large number of data inputs or an unsupervised learning method such as enhanced learning. The training process requires massive training data and a complex deep neural network structure. The huge computing amount requires ultra-high performance including computing power, precision, and scalability of processors. Nvidia GPU cluster and Google TPUs are commonly used in AI training.
  - Inferences are made using trained models and new data. For example, a video surveillance device uses the background deep neural network model to recognize a captured face. Although the calculation amount of the inference is much less than that of training, a large number of matrix operations are involved. GPU, FPGA and ASIC are also used in the inference process.

Current Status of AI Chips

Current Status of AI Chips — CPU

- **Central processing unit (CPU)**
  - The computer performance has been steadily improved based on the Moore's Law.
  - The CPU cores added for performance enhancement also increase power consumption and cost.
  - Extra instructions have been introduced and the architecture has been modified to improve AI performance.
    - Instructions, such as AVX512, have been introduced into Intel processors (CISC architecture) and vector computing modules, such as FMA, into the ALU computing module.
    - Instruction sets including Cortex A have been introduced into ARM (RISC architecture), which will be upgraded continuously.
  - Despite that boosting the processor frequency can elevate the performance, the high frequency will cause huge power consumption and overheating of the chip as the frequency reaches the ceiling.

Current Status of AI Chips — GPU

- **Graph processing unit (GPU)**
  - GPU performs remarkably in matrix computing and parallel computing and plays a key role in heterogeneous computing. It was first introduced to the AI field as an acceleration chip for deep learning. Currently, the GPU ecosystem has matured.
  - Using the GPU architecture, NVIDIA focuses on the following two aspects of deep learning:
    - **Diversifying the ecosystem:** It has launched the cuDNN optimization library for neural networks to improve usability and optimize the GPU underlying architecture.
    - **Improving customization:** It supports various data types, including int8 in addition to float32; introduces modules dedicated for deep learning. For example, the optimized architecture of Tensor cores has been introduced, such as the TensorCore of V100.
  - The existing problems include **high costs and latency** and **low energy efficiency**.

Current Status of AI Chips — TPU

- **Tensor processing unit (TPU)**
  - Since 2006, Google has sought to apply the design concept of ASICs to the neural network field and released TPU, a customized AI chip that supports TensorFlow, which is an open-source deep learning framework.
  - Massive systolic arrays and large-capacity on-chip storage are adopted to accelerate the most common convolution operations in deep neural networks.
    - Systolic arrays optimize matrix multiplication and convolution operations to elevate computing power and lower energy consumption.

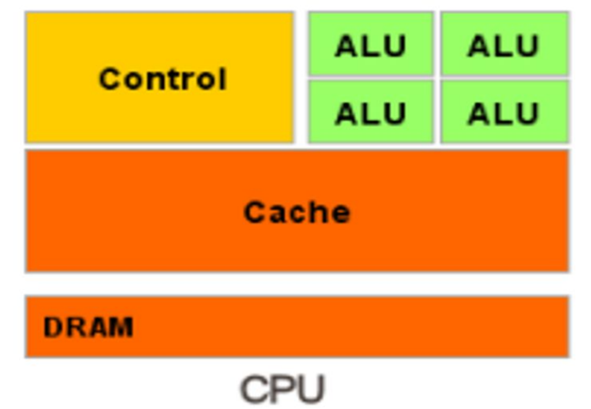
Current Status of AI Chips — FPGA

- **Field programmable gate array (FPGA)**
  - Using the HDL programmable mode, FPGAs are highly flexible, reconfigurable and re-programmable, and customizable.
  - Multiple FPGAs can be used to load the DNN model on the chips to lower computing latency. FPGAs outperform GPUs in terms of computing performance. However, the optimal performance cannot be achieved due to continuous erasing and programming. Besides, redundant transistors and cables, logic circuits with the same functions occupy a larger chip area.
  - The reconfigurable structure lowers supply and R&D risks. The cost is relatively flexible depending on the purchase quantity.
  - The design and tapeout processes are decoupled. The **development period is long**, generally half a year. **The entry barrier is high**.

Design Comparison of GPUs and CPUs

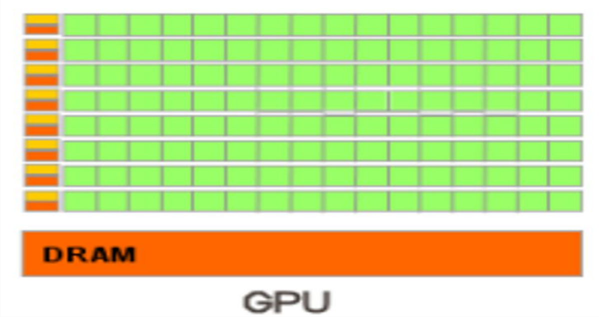
Design Comparison of GPUs and CPUs (1)

- **CPUs need to process different data types in a universal manner, perform logic judgment, and introduce massive branch jumps and interrupted processing.**
  - Composed of several cores optimized for sequential serial processing
  - Low-latency design
    - **The powerful ALU unit can complete the calculation in a short clock cycle.**
    - **The large cache lowers latency.**
    - **High clock frequency**
    - **Complex logic control unit, multi-branch programs can reduce latency through branch prediction.**
    - **For instructions that depend on the previous instruction result, the logic unit determines the location of the instructions in the pipeline to speed up data forwarding.**
  - Specialized in logic control and serial operation



Design Comparison of GPUs and CPUs(2)

- **GPUs are designed for massive data of the same type independent from each other and pure computing environments that do not need to be interrupted.**
  - Each GPU comprises several large-sized parallel computing architectures with thousands of smaller cores designed to handle multiple tasks simultaneously.
  - Throughput-oriented design
    - **With many ALUs and few caches, which improve services for threads, unlike those in CPU. The cache merges access to DRAM, causing latency.**
    - **The control unit performs combined access.**
    - **A large number of ALUs process numerous threads concurrently to cover up the latency.**
  - Specialized in computing-intensive and easy-to-parallel programs



Ascend AI Processors

Ascend AI Processors

- Neural-network processing unit (NPU): uses a deep learning instruction set to process a large number of human neurons and synapses simulated at the circuit layer. One instruction is used to process a group of neurons.
- Typical NPUs: Huawei Ascend AI chips, Cambricon chips, and IBM TrueNorth

- Ascend-Mini
- Architecture: Da Vinci
- Half precision (FP16): 8 Tera-FLOPS
- Integer precision (INT8): 16 Tera-OPS
- 16-channel full-HD video decoder: H.264/H.265
- 1-channel full-HD video decoder: H.264/H.265
- **Max. power: 8W**
- 12nm FFC

- Ascend-Max
- Architecture: Da Vinci
- Half precision (FP16): 256 Tera-FLOPS
- Integer precision (INT8): 512 Tera-OPS
- 128-channel full-HD video decoder: H.264/H.265
- Max. power: 350W
- 7nm

Thank You!

Next: 8.2 - Hardware Architecture of Ascend Chips