ICS 35.020
CCS L 70

NATIONAL STANDARD OF THE PEOPLE'S REPUBLIC OF CHINA

GB/T 45401.1-2025

Artificial intelligence - Scheduling and cooperation for computing devices - Part 1: Virtualization and scheduling

人工智能 计算设备调度与协同 第1部分：虚拟化与调度

Issued on: February 28, 2025 **Implemented on: February 28, 2025**

Issued by: State Administration for Market Regulation;
Standardization Administration of the PRC

Table of Contents

1 Scope

This document gives the architecture for the virtualization and scheduling of artificial intelligence computing devices, specifies the technical requirements and describes the test

methods.

This document applies to the system design, development and testing of the virtualization and scheduling of artificial intelligence computing devices.

2 Normative references

The contents of the following documents constitute indispensable provisions of this document through normative reference in the text. For dated references, only the edition corresponding to that date applies to this document; for undated references, the latest edition, including all amendments, applies to this document.

GB/T 41867 Information technology - Artificial intelligence - Terminology. GB/T 45087-2024 Artificial intelligence - Test methods for the performance of server systems.

3 Terms and definitions

The terms and definitions given in GB/T 41867 and the following apply to this document.

Artificial intelligence computing unit: the smallest set of components needed to carry out an artificial intelligence computing task. Note: an artificial intelligence computing unit is generally packaged in an artificial intelligence accelerator or accelerator card.

Artificial intelligence accelerating processor unit, also artificial intelligence accelerating chip: an integrated circuit component having a computing micro-architecture suited to artificial intelligence algorithms and able to carry out the computing processing of artificial intelligence applications.

Artificial intelligence accelerating card: an expansion acceleration device designed specifically for artificial intelligence computing and conforming to the hardware interface of an artificial intelligence server. Note: artificial intelligence accelerator cards are divided by scenario of use into artificial intelligence training accelerator cards, artificial intelligence inference accelerator cards and others.

Artificial intelligence computing instance: the virtualized object that carries out an artificial intelligence computing task.

Virtualization: a form of resource representation used to express decoupling from the underlying physical resources. [Source: ISO/IEC 17826:2022, 3.55]

Heterogeneous resource pool: an abstract entity formed by the collection of artificial intelligence computing resources of different architectures. Note 1: a heterogeneous resource pool provides a scalable computing architecture, which is favourable to the rational allocation of computing resources and provides guarantees of computing capability, storage, bandwidth and latency for the development and deployment of artificial intelligence application systems in different operating environments, for example cloud, cluster, mobile

device and internet of things. Note 2: a heterogeneous resource pool is used to manage and schedule artificial intelligence computing resources so as to meet the needs of different artificial intelligence computing tasks. Note 3: artificial intelligence computing resources include the central processing unit (CPU), the graphics processing unit (GPU), the neural-network processing unit (NPU), the field programmable gate array (FPGA), the digital signal processor (DSP), the application specific integrated circuit (ASIC) and others.

Computing capability: the greatest degree to which a product or system can meet a computing need.

Neural-network processing unit: an integrated circuit component specially optimized in design for neural network computation. Note: this kind of integrated circuit component is good at processing multimedia data such as video, image and speech.

Artificial intelligence computing task: the activity needed to achieve a particular artificial intelligence computing objective. Note: in this document, where no misunderstanding arises from the context, an artificial intelligence computing task generally means an inference task or a training task.

Performance: a characteristic that can be measured when a computing task is being run. Note 1: performance includes both qualitative and quantitative features. Note 2: performance is obtained from the measurement or calculation of one or more parameters, such as energy consumption, traffic, throughput, running time and rate, so as to characterize the behaviour, the features and the efficiency of a technical process running in a given machine. Note 3: in evaluating the performance of an artificial intelligence task, the throughput characteristic is generally used. [Source: ISO/IEC 20000-10:2018, 3.1.16, modified]

Artificial intelligence computing cluster: a collection of artificial intelligence computing functional units following unified control. Note 1: artificial intelligence computing functional units may include artificial intelligence accelerators, artificial intelligence servers, artificial intelligence acceleration modules and others. Note 2: when made up of artificial intelligence servers, an artificial intelligence cluster is also called an artificial intelligence server cluster.

Node: a physical or logical artificial intelligence computing device, connected by a network, that can complete a particular artificial intelligence computing task. [Source: ISO/IEC 14575:2000, 3.2.27, modified]

Scheduling: the process of controlling the place and the time at which the whole or part of a particular task is executed. Note: in this document, place generally means the artificial intelligence computing unit. [Source: ISO/IEC 10164-15:2002, 3.7.4, modified]

Scheduler: the component that carries out scheduling in a system. Note: in this document, the scheduler is used to allocate artificial intelligence computing resources for different computing needs.

Scheduling policy: the policy by which the place and the time of execution of the whole or part of a task is matched in a system so as to complete the scheduling of the task.

Isolation: the condition in which computing instances neither affect one another nor have access to one another in respect of computation and data. Note: computing power isolation means that the computing capability of one computing instance does not affect that of another. [Source: ISO/IEC TS 25052-1:2022, 3.1.5.3, modified]

4 Abbreviated terms

The following abbreviated terms apply to this document. AI, artificial intelligence. API, application programming interface. BAR, base address register. CPU, central processing unit. DDR, double data rate. DMA, direct memory access. FLOPS, floating-point operations per second. FPGA, field programmable gate array. FPS, frames per second. FP16, 16-bit floating-point number. GDDR, graphics DDR SDRAM. GPA, guest physical address. GPU, graphics processing unit. HBM, high bandwidth memory. ID, identity document. INT8, 8-bit integer number. IO, input output. IOMMU, input output memory management unit. IOVA, input output virtual address. NIC, network interface card. NPU, neural-network processing unit. OPS, operations per second. OS, operating system. PCIE, peripheral component interconnect express. QEMU, quick emulator. QoS, quality of service. SR-IOV, single root input output virtualization. VFIO, virtual function input output. VM, virtual machine. VMM, virtual machine manager. VMX, virtual machine extension.

5 Overview

The architecture for the virtualization and scheduling of AI computing devices is given in Figure 1. AI computing virtualization provides a particular form of representation for the physical AI computing resources; the virtualization schemes cover the virtualization of physical AI accelerator cards based on CPU, GPU, NPU, FPGA and the like. The several virtualization schemes form a resource pool through a unified access component, so that the physical AI computing resources are used in a consistent manner. According to the AI task and the state of the resource pool, the scheduler selects a number of virtualized AI computing instances, allocates them and executes the particular task. Operation and maintenance monitoring provides monitoring and control of the AI computing instances, of the physical AI computing resources, that is the AI accelerator cards, and of the AI tasks and their states.

Note 1: the parts shown in dashed boxes do not fall within the scope of standardization of this document. Note 2: the virtualization scheme for the FPGA covers the mixed architecture of FPGA and CPU. Note 3: one AI application is generally broken down into a number of AI computing tasks, which are handed to the scheduler.