ICS 35.240
CCS L 70

中华人民共和国国家标准

GB/T 42018—2022

信息技术　人工智能 平台计算资源规范

Information technology—Artificial intelligence—Platform computing resource specification

2022-10-12 发布　　2023-05-01 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文的发布机构不承担识别专利的责任。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：中国电子技术标准化研究院、华为技术有限公司、上海依图网络科技有限公司、杭州海康威视数字技术股份有限公司、西北工业大学、浪潮软件科技有限公司、国防科技大学、上海商汤阡誓科技有限公司、北京智芯微电子科技有限公司、北京电信规划设计院有限公司、三菱电机(中国)有限公司、青岛海尔洗衣机有限公司、英特尔(中国)有限公司、飞腾信息技术有限公司、中车沈阳机车车辆有限公司、杭州中奥科技有限公司、平安科技(深圳)有限公司、北京清能互联科技有限公司。

本文件主要起草人：董建、张琦、鲍薇、曹晓琦、李斌斌、赵春昊、周智强、符海芳、杨刚、姚远、王功明、徐洋、杨绍武、史殿习、吴庚、刘勇、毕盛楠、范科峰、马珊珊、郝守勤、马骋昊、丁晓鹏、孙宁、王海宁、夏磊、谷潇聪、殷世军、宋文林、陆韵、贾一君、李冰、杨雨泽、赵江、汪洋、江易。

引　言

人工智能平台是为人工智能应用提供各类资源的软硬件系统，是各类人工智能应用（如计算机视觉、自然语言处理、声音处理等）实现的基础。组成人工智能平台时，需使用人工智能物理计算资源和虚拟计算资源。物理计算资源是指计算设备实体，包含人工智能服务器、人工智能加速卡、人工智能加速模组等。虚拟计算资源则是基于实体计算资源，经过抽象并在一定程度上屏蔽异构性后，形成的逻辑计算资源。

本文件旨在为人工智能平台的建设提供标准依据，对可供组成人工智能平台常见的物理计算资源和虚拟计算资源提出相关技术参数的基础共性要求，对物理计算资源提出测试方法。

信息技术　人工智能
平台计算资源规范

1　范围

本文件规定了面向机器学习的人工智能平台物理计算资源(包含人工智能服务器、人工智能加速卡、人工智能加速模组)和虚拟计算资源的技术要求,描述了物理计算资源的测试方法。本文件不对资源管理和调度提出要求。

本文件适用于面向机器学习的人工智能平台的设计和测试。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB/T 9813.3—2017　计算机通用规范　第3部分:服务器

GB/T 17235.1—1998　信息技术　连续色调静态图像的数字压缩及编码　第1部分:要求和指南

GB/T 20090.2—2013　信息技术　先进音视频编码　第2部分:视频

GB/T 41867—2022　信息技术　人工智能　术语

ISO/IEC 15948:2004　信息技术　计算机图形和图像处理　便携式网络图形:功能规范[Information technology—Computer graphics and image processing—Portable Network Graphics (PNG): Functional specification]

ITU-T H.264—2021　通用视听服务高级视频编码(Advanced video coding for generic audiovisual services)

ITU-T H.265—2021　高效视频编码(High efficiency video coding)

3　术语和定义

GB/T 41867—2022界定的以及下列术语和定义适用于本文件。

3.1

人工智能平台　artificial intelligence platform

为人工智能应用提供各类资源的软硬件系统。

3.2

人工智能平台计算资源　artificial intelligence platform computing resource

在人工智能平台中,用于处理人工智能计算任务的硬件和软件。

注:人工智能平台计算资源包含物理计算资源和虚拟计算资源。

[来源:ISO/IEC 30145.3—2020,3.1.6,有修改]

3.3

物理计算资源　physical computing resource

为人工智能应用提供信息处理能力(如存储、计算等)的实体设备。

示例：人工智能服务器、人工智能加速卡和人工智能加速模组等。

3.4

虚拟计算资源　virtual computing resource

为人工智能应用提供信息处理能力(如存储、计算等)的逻辑设备。

注：逻辑设备是物理设备的虚拟化形态，它与物理设备间存在映射关系。

3.5

人工智能服务器　artificial intelligence server

信息系统中能够为人工智能应用提供高效能计算处理能力的服务器。

注 1：以通用服务器为基础，配备人工智能加速卡后，为人工智能应用提供专用计算加速能力的服务器，称人工智能兼容服务器。

注 2：专为人工智能加速计算设计，提供人工智能专用计算能力的服务器，称人工智能一体机服务器。

注 3：本文件中，在不引起误解的语境中，将人工智能服务器简称为服务器。

[来源：GB/T 41867—2022，3.1.3]

3.6

人工智能加速卡　artificial intelligence accelerating card

专为人工智能计算设计、符合人工智能服务器硬件接口的扩展加速设备。

注：本文件中，在不引起误解的语境中，将人工智能加速卡简称为加速卡。

3.7

人工智能加速模组　artificial intelligence accelerating module

专为固定领域人工智能计算设计，部署在边缘计算场景中的扩展加速部件。

注 1：人工智能加速模组一般用于执行智能摄像机、机器人、无人机等设备的人工智能计算任务。

注 2：本文件中，在不引起误解的语境中，将人工智能加速模组简称为加速模组。

[来源：GB/T 41867—2022，3.1.6]

3.8

人工智能加速处理器　artificial intelligence accelerating processor

具备适配人工智能算法的运算微架构，能够完成人工智能应用加速运算处理的集成电路元件。

[来源：GB/T 41867—2022，3.1.5，有修改]

3.9

片上系统　system on chip

大部分功能都集成在一个电路上的至少包括处理器、随机存取存储器和只读存储器的嵌入式系统。

[来源：ISO/IEC TR 18015—2006，3.45，有修改]

3.10

人工智能加速电路　artificial intelligence accelerating circuit

用来加速人工智能运算的电路。

注：本文件中，在不引起误解的语境中，将人工智能加速电路简称为加速电路。

3.11

训练　training

利用数据，基于机器学习算法，建立或改进机器学习模型参数的过程。

[来源：ISO/IEC 22989:2021，3.3.15]

3.12

推理　inference

计算机根据已知信息进行分析、分类或诊断，做出假设，解决问题或者给出推断的过程。

注：人工智能领域的推理包括逻辑推理、机器学习推理等。

[来源：GB/T 5271.28—2001，28.01.11，有修改]

4　缩略语

下列缩略语适用于本文件。

ARM：高级精简指令集处理机（Advanced Reduced Instruction Set Computer Machines）

API：应用编程接口（Application Programming Interface）

BMC：基板管理控制器（Baseboard Management Controller）

BTB：板对板（Board To Board）

CPU：中央处理单元（Central Processing Unit）

DDR3：双数据率 3（Double-Data-Rate Three）

DDR4：双数据率 4（Double-Data-Rate Four）

ECC：错误纠正码（Error Correcting Code）

FHD：全高清（Full High Definition）

FP16：16 位半精度浮点数（16 bits half-precision Floating Point）

FP32：32 位单精度浮点数（32 bits single-precision Floating Point）

FPS：帧每秒 （Frames Per Second）

GE：千兆比特以太网 （Gigabit Ethernet）

GPU：图形处理单元（Graphic Processing Unit）

HBM：高带宽内存（High Bandwidth Memory）

I2C：集成电路互联总线（Inter-Integrated Circuit）

INT4：4 位八分之一精度整型 （4 bits one-eighth-precision INTeter）

INT8：8 位四分之一精度整型（8 bits quarter-precision INTeger）

LPDDR4：低功率双数据率（Low Power Double Data Rate 4）

LPDDR4X：低功率双数据率 4 扩展（Low Power Double Data Rate 4 eXtended）

NPU：神经网络处理单元（Neural-network Processing Unit）

NVMe：非易失性内存高速通道（Non-Volatile Memory Express）

PCIe：外设部件互联高速通道（Peripheral Component Interconnect Express）

RAID：独立磁盘冗余阵列（Redundant Array of Independent Disks）

RGMII：精简千兆比特媒体独立接口（Reduced Gigabit Media Independent Interface）

SATA：串行高级连接技术（Serial Advanced Technology Attachment）

SOC：片上系统（System On Chip）

SPI：串行外设接口（Serial Peripheral Interface）

SSD：固态盘（Solid State Disk）

TCP：传输控制协议（Transmission Control Protocol）

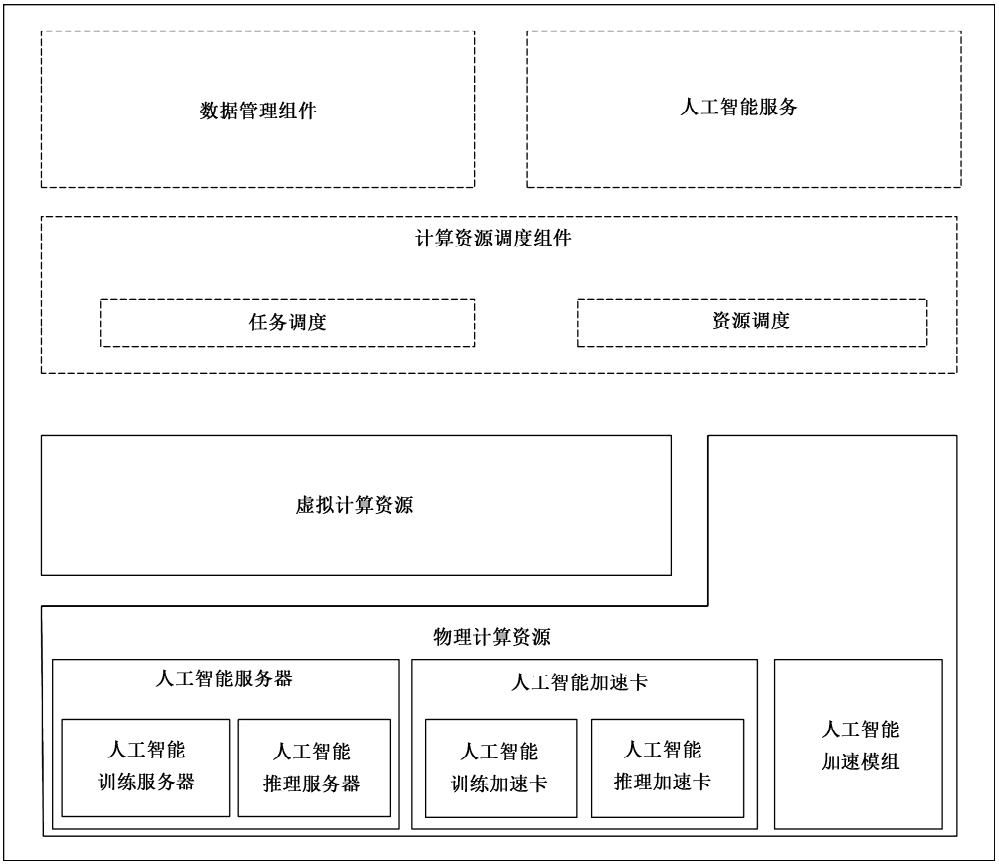
TFLOPS：每秒一万亿次浮点运算（Tera FLoating-point Operations Per Second）

TOPS：每秒一万亿次运算 （Tera Operations Per Second）

UART： 通用异步收发传输器(Universal Asynchronous Receiver Transmitter)
USB:通用串行总线(Universal Serial Bus)

5 概述

本文件描述的人工智能平台计算资源包括物理计算资源和虚拟计算资源,物理计算资源是指计算设备实体,包含人工智能服务器、人工智能加速卡、人工智能加速模组等。人工智能平台参考架构见图1。在提供人工智能服务时,通常需要管理及调度这些资源,对资源管理和调度的要求不在本文件规定的范围内。



注：图中实线框起的部分(整图外框除外)对应本文件规定的范畴,虚线框起的部分仅为表明人工智能平台的参考架构,不属于本文件规定的范畴。

图1 人工智能平台参考架构

人工智能服务器,因其应用目标和构成的差异,分为人工智能训练服务器和人工智能推理服务器。相应的技术要求和测试方法见6.1.1.1和7.2.1,及6.1.1.2和7.2.2。对同时具备训练和推理功能的人工智能服务器,在考察训练或推理相关要求时,分别适用人工智能训练服务器或人工智能推理服务器要求和测试方法。人工智能加速卡包含训练加速卡和推理加速卡,相关技术要求和测试方法见6.1.2.1和7.2.3,及6.1.2.2和7.2.4。常见的人工智能加速模组用于推理场景,本文件仅对人工智能加速模组提出面向推理的技术要求和测试方法,见6.1.3和7.2.5。本文件提出的技术要求和测试方法对应关系见附录A。

虚拟计算资源是基于实体计算资源,经过抽象并在一定程度上屏蔽异构性后,形成的逻辑计算资

源。相关基础技术要求见6.2。虚拟计算资源的测试方法由其他标准给出，本文件不做规定。虚拟计算资源不是人工智能系统实现的必要部分，计算资源的调度、人工智能服务等，可直接由物理计算资源提供。

图1是人工智能平台参考架构，实际实现时，可按需选取搭配组件。

6 技术要求

6.1 物理计算资源

6.1.1 人工智能服务器

6.1.1.1 人工智能训练服务器

对用作人工智能平台训练资源的人工智能服务器的要求如下。

a) 应符合GB/T 9813.3—2017中4.4～4.11的规定。

b) 应能提供人工智能计算加速，方式包含但不限于：

1) 通过扩展设备，如人工智能加速卡等；

2) 通过CPU加速或配套的扩展加速模组，如人工智能加速电路等；

3) 通过集成人工智能加速处理器的方式，如SOC等。

c) 应配备CPU处理核心。

d) 宜配备32位或64位多核x86或ARM CPU。

e) 宜能收集CPU运行状态。

f) 配备加速器片上内存时，总位宽宜不低于1 024 bit。

g) 应支持内存扩展(DDR3或以上版本)。

h) 一级指令缓存宜不小于64 KB，一级数据缓存宜不小于64 KB，二级缓存宜不小于512 KB，三级缓存宜不小于8 MB。

i) 宜支持DDR4及以上版本的内存，DDR控制器宜不少于4个。

j) 宜满足以下外部存储的使用要求：

1) 支持SSD、SATA或NVMe外部存储协议；

2) 支持RAID0或以上级别的可靠存储协议。

k) 宜支持PCIe 4.0协议，宜配备至少2个PCIe控制器。

l) 宜支持USB 3.0通信，配备接口。

m) 宜能连接及使用100 GE、25 GE、10 GE、GE接口。

n) 宜能完成ECC 1 bit纠错，ECC 2 bit报错。

o) 物理网口数量宜不少于4个。

p) 宜具备图像、视频预处理模块。

q) 宜能通过自带固件，如通过BMC等，或其他外接部件监控系统参数。

r) 应能执行以下至少一种场景模型的训练，包含但不限于：

1) 计算机视觉；

2) 自然语言处理；

3) 声音处理。

s) 宜支持r)1)、r)2)和r)3)场景模型的训练。

t) 应能以下列至少1种精度实施的训练：

1) INT4；
2) INT8；
3) INT16；
4) FP16；
5) FP32。

u) 通过集成人工智能加速处理器的方式提供人工智能计算加速时：
1) 宜提供不小于 16 TFLOPS(或 16 TOPS)计算能力；
2) 宜支持人工智能加速器直出参数面网口或芯片间互联。

6.1.1.2 人工智能推理服务器

对用作人工智能平台推理资源的人工智能服务器的要求如下。

a) 应符合 6.1.1.1 中 a)、b)、c)的规定。
b) 宜配备 32 位或 64 位多核 x86 或 ARM CPU。
c) 宜配备三级缓存，容量不宜低于 16 MB。
d) 宜支持 DDR4 及以上版本的内存。
e) 宜兼容 PCIe 4.0 及更低版本的 PCIe 协议。
f) 宜能连接并使用 25 GE、10 GE、GE 等接口。
g) 宜能通过自带固件，如通过 BMC 等，或其他外接部件监控系统参数。
h) 宜具备图像、视频预处理模块。
i) 应能执行以下至少一种场景模型的推理，包含但不限于：
1) 计算机视觉；
2) 自然语言处理；
3) 声音处理。
j) 宜支持 i)1)、i)2)和 i)3)场景模型的推理。
k) 应支持下列一种或多种精度实施推理：
1) INT4；
2) INT8；
3) INT16；
4) FP16；
5) FP32。

6.1.2 人工智能加速卡

6.1.2.1 人工智能训练加速卡

对用作人工智能平台训练资源的训练加速卡的要求如下。

a) 当具备卡上内存时，卡上内存带宽应不低于 100 GB/s。
b) 应至少支持乘加运算加速。
c) 配备片上内存(如 HBM)时，内存总容量应不小于 16 GB，宜不小于 32 GB，总带宽应不小于 512 GB/s，宜不小于 1 024 GB/s。
d) 应兼容 PCIe3.0 协议(PCIe x16、PCIe x8 或 PCIe x4)。
e) 宜兼容 PCIe4.0 协议。
f) 宜支持 DDR4 内存或 LPDDR4 及以上版本内存，容量宜不小于 16 GB。
g) 宜支持(如基于 ECC 的)内存错误修复。
h) 宜能解码 16 路 4 K (或 64 路 FHD)60 FPS 的视频(视频格式符合 ITU-T H.264—2021、ITU-

T H.265—2021 或 GB/T 20090.2—2013 的规定)。

i) 宜具备对 GB/T 17235.1—1998 规定格式的图像(FHD 2 048 FPS 或等效)的解码能力。

j) 宜具备对 ISO/IEC 15948:2004 规定格式的图像的解码能力。

k) 宜具备对 GB/T 17235.1—1998 规定格式的图像(FHD 256 FPS 或等效)的编码能力。

l) 应能与至少一种深度学习或机器学习框架配套运行。

m) 应能执行以下至少一种场景模型的训练,包含但不限于:

1) 计算机视觉;

2) 自然语言处理;

3) 声音处理。

n) 宜支持 m)1)、m)2)和 m)3)场景模型的训练。

6.1.2.2 人工智能推理加速卡

对用作人工智能平台推理资源的推理加速卡的要求如下。

a) 当具备卡上内存时,卡上内存带宽应不低于 32 GB/s。

b) 符合 6.1.2.1 中 b)和 d)的要求。

c) 宜支持 DDR4 或 LPDDR4 及以上版本内存,容量宜不小于 16 GB;单内存位宽不宜低于 128 bit,总带宽宜不小于 204 GB/s。

d) 宜支持(如基于 ECC 的)内存错误修复。

e) 宜支持视频编解码(视频格式符合 ITU-T H.264—2021、ITU-T H.265—2021 或 GB/T 20090.2—2013 的规定)。

f) 宜支持至少一种图像的解码,如 GB/T 17235.1—1998 或 ISO/IEC 15948:2004 规定的图像等。

g) 宜支持至少 64 路 1 080 P(即分辨率为 1 920×1 080,逐行扫描)视频解码(30 FPS)或 8 路 4 K 视频解码(60 FPS),视频格式应符合 ITU-T H.264—2021、ITU-T H.265—2021 或 GB/T 20090.2—2013 的规定。

h) 宜能支持 4 路 ISO/IEC 15948:2004 规定格式的图像[1 080 P(256 FPS)]的解码。

i) 宜能支持 4 路 ISO/IEC 15948:2004 规定格式的图像[1 080 P(64 FPS)]的编码。

j) 宜能支持 4 路 ISO/IEC 15948:2004 规定格式的图像[1 080 P(24 FPS)]的解码。

k) 应能执行以下至少一种场景模型的推理,包含但不限于:

1) 计算机视觉;

2) 自然语言处理;

3) 声音处理。

l) 宜支持 k)1)、k)2)和 k)3)中全部类型场景模型的推理。

6.1.3 人工智能加速模组

加速模组作为人工智能平台推理资源,符合以下要求。

a) 应含有至少 1 个人工智能加速处理器,支持乘加运算加速。

b) 应支持至少一种计算精度,如 FP16、INT8 等。

c) 应配备连接器或接口,如 USB、BTB 连接器等。

d) 宜能以 FP16 或 INT8 精度执行计算任务。

e) 宜兼容 LPDDR4X 协议,位宽不宜小于 64 bit,容量不宜小于 4 GB。

f) 宜支持(如基于 ECC 的)内存错误修复。
g) 宜兼容至少一种高速接口(PCIe 3.0 或以上规格接口,RGMII 或 USB 2.0 等)。
h) 宜兼容至少一种串行总线接口(UART 接口、I2C 接口或 SPI 接口等)。
i) 在高清视频解析和处理场景下,宜支持视频编解码,视频格式符合 ITU-T H.264—2021、ITU-T H.265—2021 或 GB/T 20090.2—2013 的规定:
   1) 20 路 1 080 P,25 FPS 解码或 16 路 1 080 P,30 FPS 解码;
   2) 2 路 4 K(即分辨率为 3 840×2 160),30 FPS 解码;
   3) 1 路 1 080 P,30 FPS 编码。
j) 在高清视频解析和处理场景下,宜支持下列图像的编解码:
   1) 对符合 GB/T 17235.1—1998 规定格式的图像:1 080 P,256 FPS 解码,1 080 P,64 FPS 编码;
   2) 对符合 ISO/IEC 15948:2004 规定格式的图像:1 080 P,24 FPS 解码。
k) 无源供电时,平均功耗宜不超过 5 W。
l) 应能执行以下至少一种场景模型的推理,包含但不限于:
   1) 计算机视觉;
   2) 自然语言处理;
   3) 声音处理。
m) 宜支持 l)1)、l)2)和 l)3)中全部类型场景模型的推理。

6.2 虚拟计算资源

虚拟计算资源要求如下。

a) 应具备虚拟化的 CPU。
b) 应具备一种以上虚拟化的人工智能加速处理器,如 NPU、GPU 等。
c) 应能实时监控资源状态。
d) 宜能以基于容器的方式,管理异构资源。
e) 宜支持资源池内 CPU 和人工智能加速处理器间的不同配比。
f) 宜支持基于角色的权限访问控制。
g) 宜能自动发现和维护计算资源。
h) 宜支持服务器系统与其运行应用的绑定。
i) 宜提供资源故障告警、检测和还原的功能。
j) 宜能使用硬件节能功能,包括资源回收、关闭和休眠。
k) 宜支持资源注册、使用、配额管理和操作审计。
l) 宜支持虚拟 CPU 和虚拟人工智能加速处理器的互操作。
m) 应支持至少一种深度学习或机器学习框架。
n) 应能执行以下至少一种场景模型的推理和训练,包含但不限于:
   1) 计算机视觉;
   2) 自然语言处理;
   3) 声音处理。
o) 宜能训练和推理 n)1)、n)2)和 n)3)中全部类型场景的模型。

7 测试方法

7.1 通用特性

对 GB/T 9813.3—2017 中 4.4～4.11 的测试，按 GB/T 9813.3—2017 中 5.4～5.11 的规定实施。

对本文件第 6 章中的要求，测试环境应符合 GB/T 9813.3—2017 中 5.1 的要求。

7.2 物理计算资源

7.2.1 人工智能训练服务器

人工智能训练服务器的测试中测试方法与技术要求的对应关系见附录 A，按以下操作实施。测试系统配置见附录 B。

a) 在人工智能服务器上安装操作系统和机器学习软件框架。

b) 利用框架提供的 API，编制脚本，完成张量初始化、张量乘加操作，输出结果。

c) 执行操作系统命令，查询输出 CPU 体系架构，输出 CPU 状态数据。

d) 执行操作系统命令，查询输出各级缓存的容量。

e) 执行操作系统命令，查询输出内存协议（如 DDR3，DDR4）、容量及带宽（在执行测试前，应安置相应内存，并记录内存容量）。

f) 执行操作系统命令，查询输出 SSD、SATA 及 NVMe 协议的支持情况（在执行测试前，应安置相应存储媒体）。执行系统命令，在目标存储媒体上创建、打开、关闭和删除文件。将 RAID 设备与人工智能训练服务器连接，配置形成 RAID0（至少包含两块存储媒体），命令训练服务器读取其存储媒体上的特定数据（测试前预先存放），写入 RAID0 阵列，用时为 TL1。完成后，仅保留 RAID0 阵列中的一块存储媒体，移除其他存储媒体，重复相同数据的读出和写入操作，用时为 TL2。比较 TL1 与 TL2 的大小，TL1 应不大于 TL2。

g) 执行操作系统命令，输出系统使用的 PCIe 协议版本。

h) 外接 USB 3.0 闪存盘，复制闪存盘内预先制备的文件（不小于 1 GB）到被测系统的存储媒体上，获得平均速率（测试时，应使用 USB3.0 配套的连接线缆）。

i) 使用 6.1.1.1 m）规定协议配套的路由器或交换机，连接被测系统及测试辅助通用服务器，从通用服务器使用 TCP 协议发送文件（不小于 10 GB）到被测系统，获得链路层平均传输速率（测试时，应使用相应的连接线缆）。

j) 在 a）的基础上，调用图像、视频预处理命令，输出结果。

k) 在 a）的基础上，应符合附录 C 的规定，为每个模型编制训练脚本（控制训练过程精度为 INT4、INT8、INT16、FP16 或 FP32），并配置数据集。

l) 在 k）完成后，由实验过程控制组件，对每个模型，发送训练开始指令，并计时，直到被测人工智能训练服务器输出训练结果模型时，结束计时，完成一次训练，训练时间为两次计时的差，记为 T_1。对于不使用 CPU 加速训练过程的人工智能服务器，关闭人工智能训练加速部件（或使用计算能力等同的通用服务器），完成一次训练，并计时，训练时间为 T_2。

m) 在 l）执行中，调用被测人工智能服务器提供的监控命令，获取 CPU 运行状态，如每个 CPU 的启用、禁用状态，利用率等。

n) 检查设备外观及构造，确认网口存在，执行操作系统命令，查询输出物理网口信息，使用网口接入本地交换机，针对附录 C 中的任一负载，完成分布式训练。

o) 在服务器固件中设置 ECC 为开启状态，检查 ECC 是否工作正常，或检查服务器生产者提供的

测试过程记录。

p) 在 l)执行中，调用系统监控功能(通过操作界面或命令行)，显示、记录或打印系统运行状态(包含如，CPU 个数、核数、利用率、冷却设备状态、存储媒体利用状态、固件工作状态、开机启动选项、网络连接状态、功率等)。

q) 检查设备标称计算能力。

7.2.2 人工智能推理服务器

人工智能推理服务器的测试中测试方法与技术要求的对应关系见附录 A，按以下操作实施。测试系统配置见附录 B。

a) 在人工智能推理服务器上安装操作系统。

b) 在推理服务器需配合机器学习框架完成推理时，在服务器上安装机器学习软件框架。

c) 执行 7.2.1 c)规定的操作。

d) 执行 7.2.1 d)规定的操作。

e) 执行 7.2.1 e)规定的操作。

f) 执行 7.2.1 g)规定的操作。

g) 使用 6.1.1.2 f)规定接口配套的路由器或交换机，执行 7.2.1 i)规定的操作。

h) 在 a)的基础上，调用图像、视频预处理命令，输出结果。

i) 在 a)的基础上，应符合附录 C 的规定，为每个模型，配置测试数据集，编译并安置推理模型(控制推理过程精度为 INT4、INT8、INT16、FP16 或 FP32)。

j) 在 i)完成后，由实验过程控制组件，对每个模型，发送开始指令，并计时，直到被测人工智能推理服务器输出全部推理结果时，结束计时，完成一轮推，推理总时间记为 T_3。对于不使用 CPU 加速推理过程的人工智能服务器，关闭人工智能加速部件(或使用计算能力等同的通用服务器)，完成一轮推理，记录用时为 T_4。

k) 在 j)执行中，调用被测人工智能服务器提供的监控命令，应能获取系统运行状态，包含但不限于：CPU 利用率，瞬时或平均功率。

7.2.3 人工智能训练加速卡

人工智能训练加速卡的测试中测试方法与技术要求的对应关系见附录 A，按以下操作实施。测试系统配置见附录 B。

a) 在装配了人工智能加速卡的人工智能服务器上，安装操作系统和机器学习软件框架。

b) 当具备卡上内存时，调用加速卡附带命令，查询加速卡内存情况，计算总带宽。

c) 利用框架提供的 API 或加速卡命令，编制脚本，完成张量初始化、张量乘加操作。

d) 执行 7.2.1 e)规定的操作。

e) 执行 7.2.1 g)规定的操作。

f) 在 a)的基础上，应符合附录 C 的规定，为每个模型，配置测试数据集，编译并安置模型。

g) 在 f)完成后，由实验过程控制组件，对每个模型，发送开始指令，并计时，直到被测人工智能训练服务器输出训练结果模型时，结束计时，完成一次训练，训练时间为两次计时的差，记为 T_5。对不使用 CPU 加速训练的人工智能服务器，关闭人工智能训练加速卡功能(或使用计算能力等同的通用服务器)，完成一次训练，并计时，训练时间为 T_6。

h) 使用路由器、交换机连接若干台通用服务器，用于向被测系统发送视频码流(视频格式符合 ITU-T H.264—2021、ITU-T H.265—2021 或 GB/T 20090.2—2013 的规定)。在不具备辅助

服务器时，宜使用本地文件，模拟视频流。被测服务器完成解码，并将解码后的文件保存在被测系统本地，并向测试过程控制组件发送结果：

1） 选用16路4 K视频时，使用4台服务器，每台发送4路；

2） 选用64路FHD视频时，使用8台服务器，每台发送8路。

i） 使用路由器、交换机连接16台服务器，向被测系统发送FHD图像数据（图像格式符合GB/T 17235.1—1998的规定），每台发送帧率不小于120 FPS。被测服务器完成解码，并将解码后的文件发送至实验过程控制组件。

j） 使用路由器、交换机连接8台服务器，向被测系统发送FHD图像数据（图像格式符合ISO/IEC 15948:2004的规定），每台发送帧率不小于25 FPS。被测服务器完成解码，并将解码后的文件发送至实验过程控制组件。

k） 使用路由器、交换机连接8台服务器，向被测系统发送FHD图像数据（图像格式符合GB/T 17235.1—1998的规定），每台发送帧率不小于30 FPS，被测服务器完成编码，并将编码后的文件发送至实验过程控制组件。

l） 设置ECC为开启状态，检查ECC是否工作正常，或检查生产者提供的测试过程记录。

7.2.4 人工智能推理加速卡

人工智能推理加速卡的测试中测试方法与技术要求的对应关系见附录A，按以下操作实施。测试系统配置见附录B。

a） 当具备卡上内存时，调用加速卡附带命令，查询加速卡内存情况，计算总带宽。

b） 利用加速卡命令，编制脚本，完成张量初始化、张量乘加操作。

c） 执行7.2.1 e)规定的操作。

d） 执行7.2.1 g)规定的操作。

e） 在a)的基础上，应符合附录C的规定，为每个模型，配置测试数据集，编译并安置推理模型。

f） 在e)完成后，由实验过程控制组件，对每个模型，发送开始指令，并计时，直到被测系统输出全部推理结果时，结束计时，完成一轮推理，总时间记为T_7。对于不使用CPU加速人工智能推理的人工智能服务器，关闭人工智能推理加速卡功能（或使用计算能力等同的通用服务器），完成一轮推理，并记录用时为T_8。

g） 使用路由器、交换机连接若干台通用服务器，用于向被测系统发送视频数据（视频格式符合ITU-T H.264—2021、ITU-T H.265—2021或GB/T 20090.2—2013的规定）。被测服务器完成解码，并将解码后的文件发送至测试过程控制组件：

1） 选用64路1 080P视频时，使用16台服务器，每台发送4路；

2） 选用8路4 K视频时，使用8台服务器，每台发送1路。

h） 使用路由器、交换机连接8台服务器，向被测系统发送1 080P图像数据（图像格式符合GB/T 17235.1—1998的规定），每台发送帧率为32 FPS。被测服务器完成解码，并将解码后的文件发送至实验过程控制组件。

i） 使用路由器、交换机连接2台服务器，向被测系统发送1 080P图像数据（图像格式符合GB/T 17235.1—1998的规定），每台发送帧率为32 FPS。被测服务器完成编码，并将编码后的文件发送至实验过程控制组件。

j） 使用路由器、交换机连接4台服务器，向被测系统发送1 080P图像数据（图像格式符合ISO/IEC 15948:2004的规定），每台发送帧率为32 FPS，被测服务器完成解码，并将解码后的文件发送至实验过程控制组件。

k) 设置 ECC 为开启状态，检查 ECC 是否工作正常，或检查生产者提供的测试过程记录。

l) 使用加速卡配套工具调用或命令，解码服务器上预置的 GB/T 17235.1—1998 或 ISO/IEC 15948:2004 规定的图像。

7.2.5 人工智能加速模组

人工智能加速模组的测试中测试方法与技术要求的对应关系见附录 A，按以下操作实施。测试系统配置见附录 B。

a) 利用加速模组命令，编制脚本，完成张量初始化、张量加乘操作(或加速模组支持的其他加速操作)，精度为 FP16 或 INT8。

b) 执行 7.2.1 e)规定的操作。

c) 执行 7.2.1 g)规定的操作。

d) 外接 UART-USB，I2C-USB 或 SPI-USB 转接器，传输预先制备的文件(不小于 1 GB)到加速模组，计算平均速率。

e) 直接连接 USB2.0 闪存盘，复制闪存盘内预先制备的文件(不小于 1 GB)到加速模组，计算平均速率。

f) 应符合附录 C 的规定，为每个模型，配置测试数据集，编译并安置推理模型。

g) 在 f)完成后，由测试过程控制组件，对每个模型发送测试数据到加速模组，直到被测系统输出全部推理结果时，完成一轮推理，计算附录 C 要求的推理准确率指标。

h) 使用路由器、交换机连接若干台通用服务器，用于向被测系统发送视频码流。被测系统完成编码或解码，并将结果文件保存在被测系统本地，并向测试过程控制组件发送结果：

1) 选用 20 路 1 080P，25 FPS 待解码视频时，使用 4 台服务器，每台发送 5 路；

2) 选用 16 路 1 080P，30 FPS 待解码视频时，使用 4 台服务器，每台发送 4 路；

3) 选用 2 路 4K，30 FPS 待解码视频时，使用 2 台服务器，每台发送 1 路；

4) 选用 1 路 1 080P，30 FPS 待编码视频时，使用 1 台服务器，发送 1 路。

i) 使用路由器、交换机连接 8 台服务器，向被测系统发送 1 080P 图像数据(图像格式符合 GB/T 17235.1—1998 的规定)，每台发送帧率为 32 FPS。被测系统完成解码，并将解码后的文件发送至实验过程控制组件。

j) 使用路由器、交换机连接 2 台服务器，向被测系统发送 1 080P 图像数据(图像格式符合 GB/T 17235.1—1998 的规定)，每台发送帧率为 32 FPS。被测系统完成编码，并将编码后的文件发送至实验过程控制组件。

k) 使用路由器、交换机连接 3 台服务器，向被测系统发送 1 080P 图像数据(图像格式符合 ISO/IEC 15948:2004 的规定)，每台发送帧率为 8 FPS。被测系统完成解码，并将解码后的文件发送至实验过程控制组件。

l) 测试主机向被测样机输入测试集数据，同步启动功率计对被测样机在执行测试期间的功耗进行测量和记录，测试主机读取功率计记录的总功耗值，并计算得到该被测样机的功率，重复多次求平均，获得平均功率。

m) 使用 RGMII 连接测试系统与加速模组，从测试系统向加速模组发送数据，并计算数据传输速率。

n) 设置 ECC 为开启状态，检查 ECC 是否工作正常，或检查生产者提供的测试过程记录。

7.3 虚拟计算资源

虚拟计算资源测试方法由产品标准给出。

附 录 A
（资料性）
人工智能平台物理计算资源测试方法与技术要求的对应关系

人工智能训练服务器的测试方法与技术要求的对应关系见表 A.1，人工智能推理服务器的测试方法与技术要求的对应关系见表 A.2，人工智能训练加速卡的测试方法与技术要求的对应关系见表 A.3，人工智能推理加速卡的测试方法与技术要求的对应关系见表 A.4，人工智能加速模组的测试方法与技术要求的对应关系见表 A.5。

表 A.1 人工智能训练服务器测试方法与技术要求对应关系

测试对象	技术要求	测试方法
通用特性	6.1.1.1 a)	7.1
中央处理器	6.1.1.1 c)	7.2.1 c)
	6.1.1.1 d)	7.2.1 c)
	6.1.1.1 h)	7.2.1 d)
人工智能加速部件	6.1.1.1 b) 6.1.1.1 t)	7.2.1 k) 7.2.1 l)
	6.1.1.1 u)1)	7.2.1 q)
内存	6.1.1.1 f)或 6.1.1.1 g)或 6.1.1.1 i)	7.2.1 e)
	6.1.1.1 n)	7.2.1 o)
PCIe	6.1.1.1 k)	7.2.1 g) 7.2.1 l)
存储媒体	6.1.1.1 j)	7.2.1 f)
互联	6.1.1.1 l)	7.2.1 h)
	6.1.1.1 m)	7.2.1 i)
	6.1.1.1 o)	7.2.1 n)
	6.1.1.1 u)2)	7.2.1 n)
数据预处理	6.1.1.1 p)	7.2.1 j)
训练	6.1.1.1 r)或 6.1.1.1 s)	7.2.1 k) 7.2.1 l)
运行监控	6.1.1.1 e)	7.2.1 m)
	6.1.1.1 q)	7.2.1 p)

表 A.2　人工智能推理服务器测试方法与技术要求对应关系

测试对象	技术要求	测试方法
通用特性	6.1.1.1 a)	7.1
中央处理器	6.1.1.1 c)	7.2.2 c)
	6.1.1.2 b)	7.2.2 c)
	6.1.1.2 c)	7.2.2 d)
人工智能加速部件	6.1.1.1 b) 6.1.1.2 k)	7.2.2 i) 7.2.2 j)
内存	6.1.1.2 d)	7.2.2 e)
PCIe	6.1.1.2 e)	7.2.2 f) 7.2.2 i) 7.2.2 j)
互联	6.1.1.2 f)	7.2.2 g)
数据预处理	6.1.1.2 h)	7.2.2 h)
推理	6.1.1.2 i)或 6.1.1.2 j)	7.2.2 i) 7.2.2 j)
运行监控	6.1.1.2 g)	7.2.2 k)

表 A.3　人工智能训练加速卡测试方法与技术要求对应关系

测试对象	技术要求	测试方法
软件框架	6.1.2.1 l) 6.1.2.1 b)	7.2.3 a) 7.2.3 c)
内存	6.1.2.1 a) 6.1.2.1 c) 6.1.2.1 f)	7.2.3 d) 7.2.3 b)
	6.1.2.1 g)	7.2.3 l)
PCIe	6.1.2.1 d) 6.1.2.1 e)	7.2.3 e) 7.2.3 f) 7.2.3 g)
编解码	6.1.2.1 h)	7.2.3 h)
	6.1.2.1 i)	7.2.3 i)
	6.1.2.1 j)	7.2.3 j)
	6.1.2.1 k)	7.2.3 k)
训练	6.1.2.1 m)或 6.1.2.1 n)	7.2.3 f) 7.2.3 g)

表 A.4　人工智能推理加速卡测试方法与技术要求对应关系

测试对象	技术要求	测试方法
加速计算	6.1.2.1 b)	7.2.4 b)
内存	6.1.2.2 c) 6.1.2.2 a)	7.2.4 a) 7.2.4 c)
	6.1.2.2 d)	7.2.4 k)
PCIe	6.1.2.1 d)	7.2.4 d)
编解码	6.1.2.2 e)或 6.1.2.2 g)	7.2.4 g)
	6.1.2.2 h)	7.2.4 h)
	6.1.2.2 i)	7.2.4 i)
	6.1.2.2 j)	7.2.4 j)
	6.1.2.2 f)	7.2.4 l)
推理	6.1.2.2 k)或 6.1.2.2 l)	7.2.4 e) 7.2.4 f)

表 A.5　人工智能加速模组测试方法对应关系

测试对象	技术要求	测试方法
加速器及加速运算	6.1.3 a)	7.2.5 a)
精度支持	6.1.3 b)或 6.1.3 d)	7.2.5 a)
连接器	6.1.3 c)	7.2.5 a)
内存	6.1.3 e)	7.2.5 b)
	6.1.3 f)	7.2.5 n)
PCIe	6.1.3 g)的 PCIe 或 以上规格接口	7.2.5 c)
RGMII	6.1.3 g)的 RGMII 接口	7.2.5 m)
USB 2.0	6.1.3 g)的 USB 2.0 接口	7.2.5 e)
串行总线接口	6.1.3 h)	7.2.5 d)
编解码	6.1.3 i)	7.2.5 h)
	6.1.3 j)1)	7.2.5 i) 7.2.5 j)
	6.1.3 j)2)	7.2.5 k)

表 A.5　人工智能加速模组测试方法对应关系（续）

测试对象	技术要求	测试方法
推理	6.1.3 l)或 6.1.3 m)	7.2.5 f) 7.2.5 g)
功率	6.1.3 k)	7.2.5 l)

附 录 B
（资料性）
物理计算资源测试系统的配置

B.1 人工智能服务器测试系统配置

人工智能服务器的测试系统配置见图 B.1。

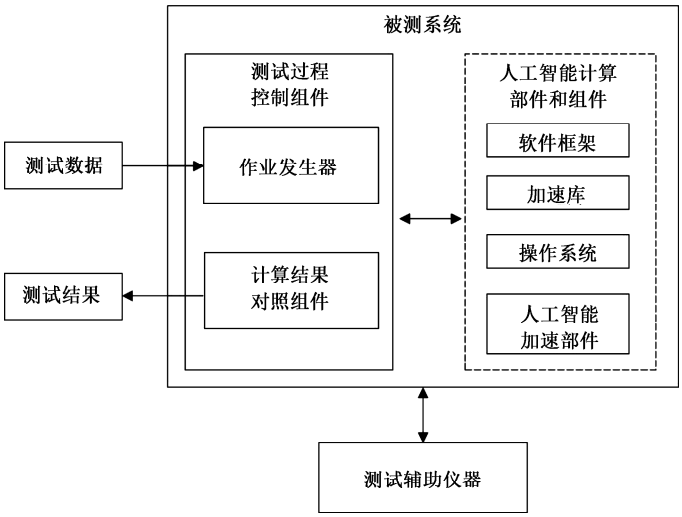


图 B.1 人工智能服务器物理计算资源的测试配置

人工智能服务器物理计算资源的测试配置说明如下：

a) 被测系统为人工智能训练服务器或人工智能推理服务器；

b) 测试过程控制组件安置、运行于被测系统内；

c) 作业发生器使用测试数据，形成人工智能作业，如使用特定数据集的训练或推理任务等，计算结果对照组件，对比返回的结果与预期值；

注：预期值指被测系统在正常工作时，正确、完整执行特定任务的结果数据或状态参数值。

d) 人工智能计算部件和组件，包含执行人工智能测试任务的加速部件，如加速卡、加速器等，和必要的软件组件，如操作系统，加速库和软件框架等；

e) 测试辅助仪器包含但不限于：闪存盘、加速卡、功率计、存储媒体（如 SSD、SATA 存储媒体等）、RAID 卡、交换机、通用服务器、必要的连接线缆和电源；

f) 测试数据，来自被测系统外部，在测试前可预先安置在被测系统内，供测试使用。

B.2 人工智能加速卡测试系统配置

人工智能加速卡的测试系统配置见图 B.2。

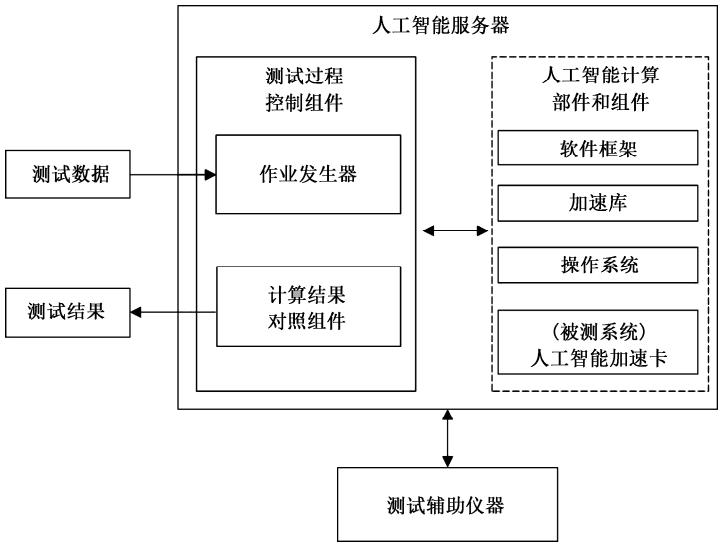


图 B.2　人工智能加速卡物理计算资源的测试配置

人工智能加速卡物理计算资源的测试配置说明如下：

a) 被测系统为人工智能加速卡，安置在人工智能服务器内。实施对特定加速卡的测试时，人工智能服务器内其他人工智能加速部件不应开启；

b) 测试过程控制组件安置、运行于服务器内；

c) 作业发生器使用测试数据，形成人工智能作业，如使用特定数据集的训练或推理任务等，计算结果对照组件，对比返回的结果与预期值；

注：预期值指被测系统在正常工作时，正确、完整执行特定任务的结果数据或状态参数值。

d) 人工智能计算部件和组件，包含执行人工智能测试任务的加速部件，如加速卡、加速器等和必要的软件组件，如操作系统，加速库和软件框架等；

e) 测试辅助仪器包含但不限于：加速卡、功率计、存储媒体（如 SSD、SATA 存储媒体等）、RAID 卡、交换机、通用服务器、必要的连接线缆和电源；

f) 测试数据，来自被测系统外部，在测试前可预先安置在被测系统内，共测试使用。

B.3　人工智能加速模组系统配置

人工智能加速模组的测试系统配置见图 B.3。

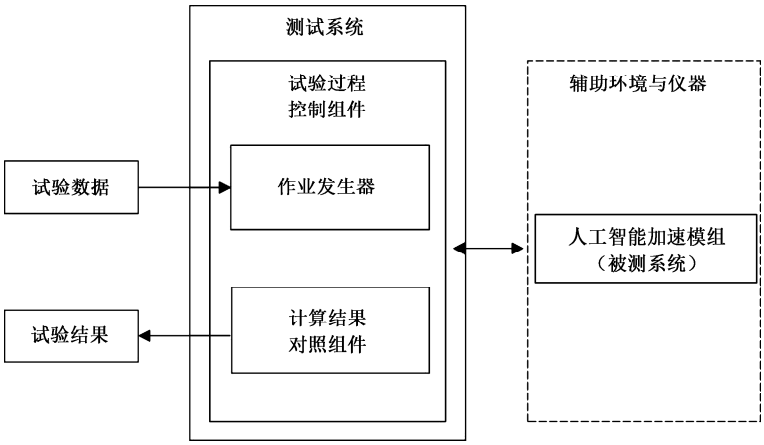


图 B.3　人工智能加速模组物理计算资源的测试配置

人工智能加速模组物理计算资源的测试配置说明如下：

a）　被测系统为人工智能加速模组，以 6.1.3 g)、6.1.3 h)或 6.1.3 c)规定的连接方式，接入测试辅助环境；

b）　被测系统测试辅助仪器包含但不限于：功率计、电源、连接器公坐、连接线缆及其他必要的能使人工智能加速模组运行并实施参数计量的软硬件；

c）　测试系统是一台或多台服务器，并配备了测试过程控制组件；

d）　测试数据，来自被测系统外部，在测试前预先安置在测试系统内，共测试使用。

附 录 C
（规范性）
训练和推理测试模型

计算机视觉训练和推理测试模型见表 C.1，自然语言处理训练和推理测试模型见表 C.2，声音处理训练和推理测试模型见表 C.3。

表 C.1 计算机视觉模型

模型	resnet50 v1.5
数据集	imagenet2012
门限	Top1 的准确率 ＞ 74%（FP32）
注：resnet 是残差神经网络。	

表 C.2 自然语言处理模型

模型	bert-large
数据集	训练：cn-wiki 或 en-wiki 推理：Squad 1.1 或 Squad 2.0
门限	训练：Mask_lm_accuracy ＞ 0.7（FP32） 推理：F1 ＞ 0.91（FP32）
注：bert-large 是基于变换器的双向编码表示的神经网络。	

表 C.3 声音处理模型

模型	FCN-4
数据集	Music-tagging
门限	Top1 的准确率 ＞ 90%（FP32）
注：FCN 是全卷积神经网络。	

参 考 文 献

[1] GB/T 5271.28—2001 信息技术 词汇 第28部分:人工智能 基本概念与专家系统

[2] ISO/IEC TR 18015:2006 Information technology—Programming languages, their environments and system software interfaces—Technical report on C++ performance

[3] ISO/IEC 19099:2014 Information technology—Virtualization management specification

[4] ISO/IEC 22989:2021 Information technology—Artificial intelligence—Artificial intelligence concepts and terminology

[5] ISO/IEC 30145.3:2020 Information technology—Smart city ICT reference framework—Part 3: Smart city engineering framework