ICS 35.040
CCS L 71

团　　体　　标　　准

T/AI 118.1—2022

人工智能算力中心
第 1 部分：技术要求

Artificial intelligence computing centre
Part 1: Technical requirements

2022－12－30 发布　　2022－12－30 实施

中关村视听产业技术创新联盟　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。

本文件是T/AI 118《人工智能算力中心》的第1部分。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由新一代人工智能产业技术创新战略联盟AI标准工作组提出。

本文件由中关村视听产业技术创新联盟归口。

本文件起草单位：鹏城实验室、清华大学、北京大学、北京市商汤科技开发有限公司、中科寒武纪科技股份有限公司、华为技术有限公司、上海燧原科技有限公司、百度在线网络技术（北京）有限公司、平安科技（深圳）有限公司。

本文件主要起草人：任智祥、陈文光、曾炜、吕文静、张鹏、赵海英、汪邦虎、张世雄、李若尘、李志永、肖京、吴庚、赵轩、黄乾明、黄岩哲、姚伟峰、侍国斌、桂煌、赵淑静、胡敏、边思雨、熊亮、陈又新。

引　言

T/AI 118《人工智能算力中心》旨在为人工智能算力中心的建设、测试等方面提供规范化指引和依据。由以下2个部分构成：

——第1部分：技术要求。目的在于确定人工智能算力中心的组成、性能、可靠性技术要求。

——第2部分：测试方法。目的在于确立人工智能算力中心性能、可靠性特性的测试方法，实现对为系统优化、瓶颈发现提供试验依据。

T/AI 118《人工智能算力中心》规定人工智能算力中心的技术特性要求和测试方法，不涉及算力中心土建、机房设计、设施安全等内容。

T/AI 118.1中5.2，5.3，6.1和7.1所提基础要求的测试，见T/AI 118.2中的5.2，5.3，附录A和附录B。

人工智能算力中心
第 1 部分：技术要求

1 范围

本文件规定了人工智能算力中心的组成、性能、可靠性技术要求。

本文件适用于人工智能算力中心的设计和建设，也为人工智能算力中心能力测试提供参考和依据。

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 41867 — 2022 信息技术 人工智能 术语

IEEE 754 — 2019 IEEE 浮点算数标准 （IEEE Standard for Floating-Point Arithmetic）

3 术语

GB/T 41867 — 2022 界定的及下列术语和定义适用于本文件。

3.1

数据中心 data centre

一种能够提供容纳、互联和操作的结构，或结构组。它使用信息技术、电信网络设备提供的数据存储、处理、迁移服务及其它所有功能，并集成能量供应、环境控制和为保证服务可用性而制定的必要的韧性、安全性级别定义。

注 1：数据中心结构一般包含数个楼宇或空间，用以支撑数据中心主要功能。

注 2：包含数据中心中信息及通信技术设备及支撑环境控制设备边界或空间，定义于更大的结构或楼宇中。

3.2

算力中心 computing centre

为多用户提供计算服务的设施。用户的操作通过对计算设备及辅助硬件的操作及中心人员的服务实现。

3.3

人工智能算力中心 artificial intelligence computing centre

一种能够为多用户提供人工智能计算服务、数据容纳的结构或结构组。使用信息技术、电信网络设备提供的数据存储、处理、迁移，人工智能计算加速等功能，并集成能量供应、环境控制和为服务可用性而制定的必要的可靠性组件。

注 1：人工智能算力中心一般包含数据中心可能涉及的楼宇或空间，用以支撑人工智能算力中心主要功能。

注 2：人工智能算力中心中的服务器，一般包含人工智能服务器和通用服务器等，服务器称为“节点”。

3.4

性能 performance

运行计算任务时，可被测量的特性。

注：性能可基于一个或多个参数（如运行时间、能耗、吞吐率、有效计算能力、每秒浮点运算次数等）的测量或计算获得，以表征在某设备（组）中运行的某技术过程的行为、特性及效率。

3.5

可靠性 reliability

实施一致的期望行为并获得结果的性质。

3.6

[工作]负载 [work]load

为测试目的，运行在计算系统中的给定任务集合。

注：一般包含输入、输出要求，计算数量和种类及所要求的计算资源。

3.7

有效计算能力 effective computing ability

在给定任务集合上，对每个任务的实际吞吐率与其基线吞吐率之比的加权几何平均。

3.8

每秒浮点运算次数 floating point operations per second

在执行某项任务过程中，关于特定种类操作的每秒执行浮点运算次数。

注 1：对训练、推理任务，一般设定特定种类操作的范围（如模型上的前向或反向计算所涉及的操作）。

注 2：如执行训练、推理任务的计算设备使用整型（如 INT16）数据类型，则称“每秒整型运算次数”。

4 缩略语

下列缩略语适用于本文件。

AUTOML 自动机器学习 （Automatic Machine Learning）
BMC 基板管理控制器 （Baseboard Mangement Controller）
CE 可改正错误 （Correctable Error）
COP 性能系数（Coefficient Of Performance）
CPU 中央处理单元（Central Processing Unit）
CVE 通用漏洞披露 （Common Vulnerabilities and Exposures）
ECC 错误纠正码 （Error-Correcting Code）
EOR 行末交换机 （End Of Row）
GE 千兆以太网或吉咖比特以太网 （Gigabit Ethernet）
HBM 高带宽存储器 （High Bandwidth Memory）
DCMI 数据中心管理接口（Data Center Manageability Interface）
DDR 双倍数据率 SDRAM （Double Data Rate SDRAM）
IO 输入输出 （Input Output）
LACP 链路聚合控制协议 （Link Aggregation Control Protocol）
MLAG 跨设备链路聚合组 （Multichassis Link Aggregation Group）

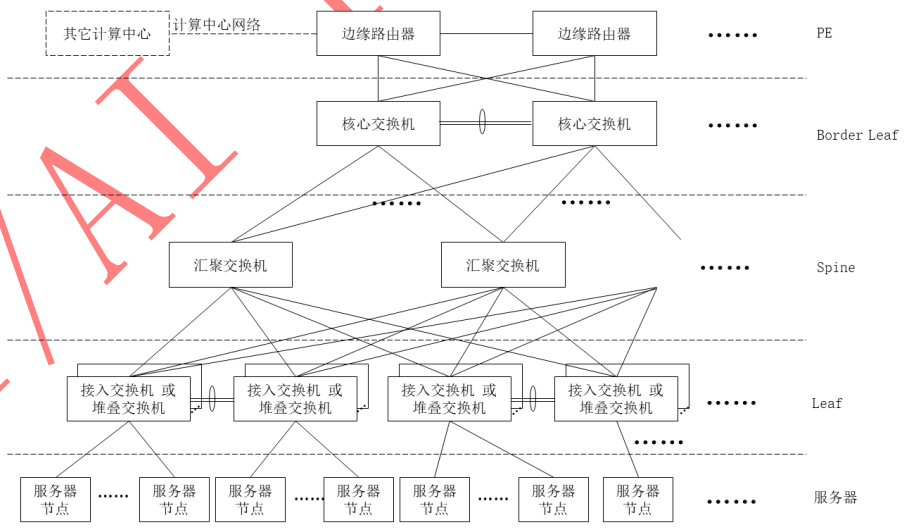
NCSI 网络控制器边带接口 （Network Controller Sideband Interface）
NVMe 非易失性存储器标准 （Non-Volatile Memory Express）
PCIE 外设部件互联高速通道（Peripheral Component Interconnect Express）
PE 提供者边缘 （Provider Edge）
SAS 串行 SCSI（Serial Attached SCSI）
SATA 串行高级技术附件（Serial Advanced Technology Attachment）
SCSI 小型计算机系统接口 （Small Computer System Interface）
SDI 数字分量串行接口 （Serial Digital Interface）
SDRAM 同步动态随机存储器 （Synchronous Dynamic Random Access Memory）
SSD 固态硬盘（Solid State Disk）
RAID 独立磁盘冗余阵列 （Redundant Arrays of Independent Drives）
RELU 修正线性单元（Rectified Linear Uint）
TOR 架顶交换机 （Top Of Rack）
UEFI 统一的可扩展固件接口 （Unified Extensible Firmware Interface）
USB 通用串行总线（Universal Serial Bus）
VRRP 虚拟路由冗余协议 （Virtual Router Redundancy Protocol）
YAM YARN 应用主导（YARN Application Master）
YARN 另一种资源协调者（Yet Another Resource Negotiator）

5 概述

5.1 参考架构

5.1.1 硬件及互联

人工智能算力中心节点及网络连接参考架构见图 1。

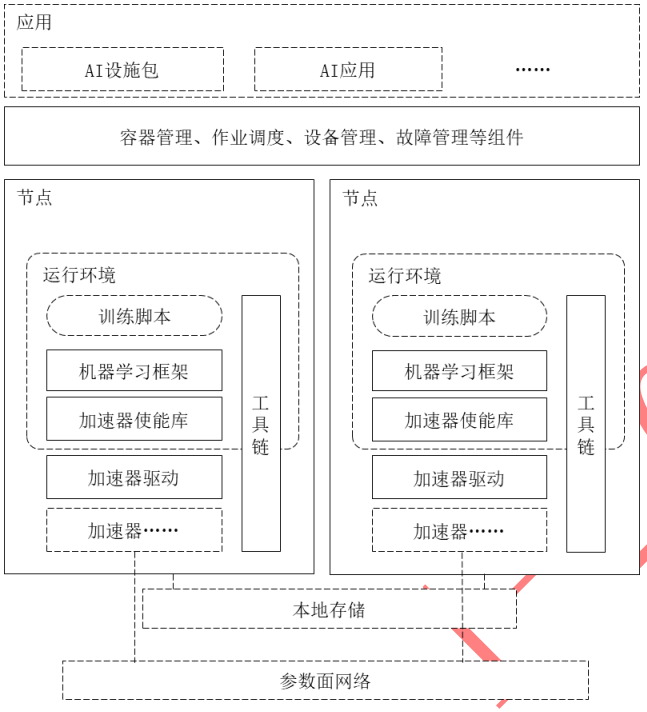


注 1：图中 PE 层中，用虚线框表示的部分不属于人工智能算力中心内的连接参考架构，本文件不对其作出规定。
注 2：本图为算力中心内部连接参考架构，各算力中心实际实现可与此有差异。

图 1 人工智能算力中心节点及网络参考架构（三级交换架构）

5.1.2 软件

人工智能算力中心的软件组件是用户应用使用算力中心功能的媒介，应参考图 2 设计和部署：



注 1：图中用虚线框表示硬件设备、作业或不属于人工智能算力中心必须具备的软件组件。

注 2：图中的运行环境一般指容器或其他能支撑节点内软件运行的必要组件。

图 2 人工智能算力中心软件组件架构参考

5.2 组成要求

人工智能算力中心部件，应符合以下规定：

a） 具备专为加速人工智能计算而设计、实现的计算部件，包含但不限于：

1） 人工智能加速处理器；

2） 人工智能服务器。

b） 具备能加速人工智能计算的互联部件，包含但不限于：

1） 人工智能加速处理器片上互联控制模组和连接机构；

2） 人工智能加速处理器机内互联控制模组和连接机构；

3） 人工智能服务器机间互联控制模组和连接机构。

5.3 供应链要求

为保障人工智能算力中心的建设和运行，选备、优化供应链，符合以下要求：

a） 应具备完整的硬件系统供应链及实践成功记录，包含但不限于供应商寻源，采购，生产，仓储，物流网络，销售，售后服务等；

b） 人工智能算力中心使用的人工智能加速设备的提供者宜具备以下部件的独立研制能力，包含但不限于：

1） 人工智能核心加速部件（包含加速器、加速板/卡）；

2） 服务器间连接组件（包含高速连接协议、线缆、端子及配套软件）；

3) 服务器配套的软件组件（包含算子实现库、机器学习框架）；
4) 服务器配套的应用开发接口（至少包含一个技术领域，如计算机视觉）；
5) 服务器配套的应用开发环境和相关工具（包含集成开发环境、调试工具、编译器、部署工具、模型性能自动优化搜索工具）；
6) 服务器配套的存储、网络通讯设备（包含存储服务器、交换机、路由器），能与服务器配套形成集群；

c） 机器学习框架宜不基于其它上游框架项目研制，其演进也宜不依赖于其它上游框架；
d） 具备所交付人工智能算力中心中设备的配套运维团队和工具。

6 性能要求

6.1 基础要求

6.1.1 训练服务器

人工智能算力中心使用的人工智能训练服务器，符合以下要求：
a） 应支持至少两种深度学习或机器学习框架；
b） 应支持 DDR4 或以上版本的内存，宜支持不少于 4 个 DDR 控制器；
c） 应支持 SAS、SATA 或 NVMe 等存储协议；
d） 宜支持计算机视觉，自然语言处理，声音处理，强化学习和自动驾驶场景模型的训练；
e） 应支持 PCIE 协议，版本不低于 3.0，宜支持至少 2 个 PCIE 控制器；
f） 应支持 USB 2.0 通信，配备接口；
g） 应支持 100GE、25GE、10GE、GE 接口
h） 支持片间数据通道或接口，单向通信速率不低于 24Gbps；
i） 宜支持人工智能加速处理器芯片直出的参数面网口；
j） 采用人工智能加速器片上内存时，片上内存不宜低于 32GB，总带宽不宜小于 1200GB/s；
k） 采用板载内存时，板载内存不宜低于 48GB；
l） 应支持图像、视频预处理；
m） 服务器整机电源功率不应低于 2KW；
n） 单条内存容量应不小于 32GB，宜能支持 64GB 或以上单条存容量；
o） 可配内存数量应不低于 16 条，宜不低于 32 条。

6.1.2 推理服务器

人工智能算力中心使用的人工智能推理服务器，符合以下要求：
a) 应支持至少 1 个独立或集成的 CPU；
b) 宜支持 L3 缓存，容量不低于 16MB；
c) 宜支持 DDR4 或以上版本的内存；
d) 应支持 PCIE 协议，版本不低于 3.0；
e) 应支持 25GE、10GE、GE 等网络接口；
f) 应支持图像、视频预处理模块；
g) 应支持计算机视觉，自然语言处理，声音处理场景模型推理；
h) 应支持 SAS、SATA 或 NVMe 等存储协议；
i) 服务器整机电源功率不应低于 500W；
j) 单条内存容量不应小于 16GB，宜能支持 64GB 或以上单条存容量；

k） 可配内存数量应不低于 8 条，宜不低于 24 条。

6.2 扩展测试负载

6.2.1 固定负载

人工智能算力中心的性能评价，如使用固定负载，应包含但不限于表 1 规定的模型及相关设定：

表 1 训练、推理固定负载

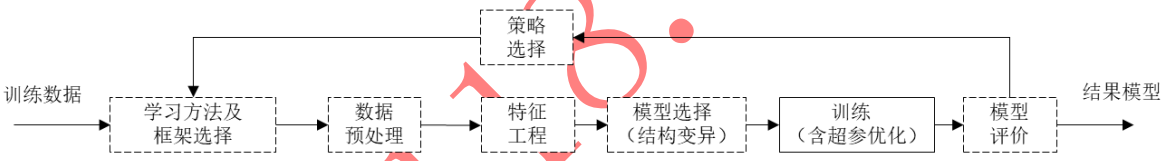
类型	项目	技术要素	人工智能算力中心
A. 图像识别	1	模型	resnet101_v1
		数据集[a]	imagenet2012
		门限[b]	Top1-准确率 > 75%
		优化方法	sgd+momentum
		试验次数	7
		结果模型精度[e]	FP16 / FP32 [f]
		损失函数	softmax + cross entropy loss
	2	模型	resnet50_v1.5
		数据集[a]	imagenet2012
		门限[b]	Top1-准确率 > 74%
		优化方法	sgd + momentum
		试验次数	7
		结果模型精度[e]	FP16 / FP32 [f]
		损失函数	softmax + cross entropy loss
B. 自然语言处理	1	模型	bert-large [c d]
		数据集[a]	训练：cn-wiki / en-wiki；推理：SQuAD v1.1
		门限[b]	训练：mask_lm_accuracy > 0.7 推理：F1 > 91.0
		优化方法	Lamb
		试验次数	7
		结果模型精度[e]	FP16 / FP32 [f]
		损失函数	softmax + negative maximum likehood loss
	2	模型	盘古-alpha
		试验次数	1
		数据集	Chinese text corpus[3]
		门限	零次学习 F1 (CMRC2018)： 16.647 零次学习 F1 (DRCD) ： 9.99 零次学习 F1 (WebQA)： 16.32
		优化方法	Adam

表 1 训练、推理固定负载 （续）

类型	项目	技术要素	人工智能算力中心
B. 自然语言处理	2	结果模型精度[e]	FP16 / FP32[f]
		损失函数	softmax + negative maximum likehood loss
注：表中的“/”符号表示“或”。			
[a] 训练数据的格式，没有统一限定，被测者可以根据本地系统组成实施必要的格式转换，格式转换过程不应改变数据的值（如图像像素值），数据格式转换过程不计时。 [b] 表中门限为参考值，测试实施时可作调整，但应在各被测系统受测时保持统一。 [c] bert-large 测试项中，sequence-length 在训练时取值为 512，推理时取值为 384。 [d] 表中 bert-large 的损失函数定义与实现见[1]。 [e] 在推理时，使用 INT8 或 FP16。 [f] 浮点算数格式宜符合 IEEE 754—2019 的要求。			

6.2.2 可变负载

人工智能算力中心的性能评价，如使用基于AUTOML[4]的可变负载（测试项目号记为C）（参考[2]），可变负载（见图3）仅含有对中间构型的模型的训练过程，此过程包含按既定策略的超参变化过程；



注：图中的虚线框不包含在可变负载的测试计时、计能等过程中。

图 3 自动机器学习参考流程

a） 可变负载含有种子模型，种子模型及中间过程模型所使用的算子，包含但不限于：

1） 卷积层（convolutional layer）；
2） 全连接层（fully connected layer）；
3） 批归一化层（batch normalization）；
4） 修正线性单元 （RELU）；
5） 加法层 （addition layer）；
6） 最大池化层（max-pooling layer）；
7） 全局池化层 （global-pooling layer）；
8） 归一化指数函数 （softmax layer）；

b） 变负载的模型结构变异指新增、减少 a）提出的结构，并可更改连接。

6.3 扩展指标

6.3.1 时间

6.3.1.1 训练总用时

训练总用时 T_T（ms）是从训练开始读入数据的时点 t_{T1}，到模型训练完毕、完成在非电易失性存储器上的持久化的时点 t_{T2}，所使用的总时长，见式 1。

$$T_T = t_{T2} - t_{T1} \quad (1)$$

6.3.1.2 训练启动用时

训练启动用时 T_W（ms）是多加速器训练时，从训练开始命令调用开始发出时点 t_{W1}，到所有加速器都被分配并开始执行训练任务的时点 t_{W2}所经历的时长，见式 2。

注 1：所有加速器都开始执行训练任务的时点是最后一个加速器开始执行训练任务的时点。

注 2：训练启动用时不包含数据读取用时。

$$T_W = t_{W2} - t_{W1} \quad (2)$$

6.3.1.3 训练用时

训练用时 T_{TR}（ms）是从训练开始命令调用开始时点 t_{TR1}（$t_{TR1} = t_{W1}$）到训练退出时点 t_{TR2}之间的时间间隔，见式 3。

注 1：训练退出可有多种充分条件。正常结束训练退出条件，如测试集准确率门限等。

注 2：训练用时不包含模型持久化用时。

$$T_{TR} = t_{TR2} - t_{TR1} \quad (3)$$

6.3.1.4 推理总延时

推理总延时 T_I（ms）是多次连续推理端到端总延时，是测试者发送第一个样本的时点 t_{I1}，到接收到最后一个样本推理结果的返回的时点 t_{I2}的差，见式 4。

注：其中每一次推理，包含必要的数据预处理，后处理过程。

$$T_I = t_{I2} - t_{I1} \quad (4)$$

6.3.1.5 端到端推理延时

端到端推理延时 T_{TI}（ms）是测试者发送样本时点 t_{TI1}与测试者收到样本的推理结果时点 t_{TI2}的差，见式 5。

注：端到端推理延时，包含必要的数据预处理，后处理过程。

$$T_{TI} = t_{TI2} - t_{TI1} \quad (5)$$

6.3.1.6 分派处理延时

分派处理延时T_{DIP}（ms）被测者完整收到样本的时点t_{DIP1}与处理结束时间t_{DIP2}的差，见式6。

$$T_{DIP} = t_{DIP2} - t_{DIP1} \quad (6)$$

6.3.2 吞吐率

6.3.2.1 训练吞吐率

针对特定训练任务，人工智能算力中心（含 N 个实际执行训练任务的节点）在训练过程中，训期（epoch）中处理的样本个数与训期平均用时（T_{EP}）的比值，见式 7。

$$Th_T = \sum_{n=1}^{N} \frac{训练集样本数量}{T_{EP-n}} \quad (7)$$

6.3.2.2 推理吞吐率

人工智能算力中心在单位时间内，对特定任务负载，完整处理，成功返回结果的样本数量。对于计算机视觉类的任务，单位是“处理的图像张数/s”；对于自然语言处理任务，单位是“句数/s”，见式 8。

$$Th_I = \frac{成功返回结果的样本数量}{T_I} \quad (8)$$

6.3.3 有效计算能力

人工智能算力中心在给定任务集合 S 上，每任务 s 的实际吞吐率Th_s与基线吞吐率Th_s^*之比的加权几何平均（参考[5]），见式 9。

注 1：使用某特定参照计算系统，在特定负载上测得的训练或推理吞吐率。

注 2：有效计算能力适用于训练和推理测试。

$$\overline{Th} = \alpha \cdot \sqrt[\sum \tau_s]{\prod_s \left(\frac{Th_s}{Th_s^*}\right)^{\tau_s}} \quad (9)$$

式中：

α —— 调整系数（$\alpha > 0$，$\alpha \in \mathbb{R}^+$），默认为 100.0；

s —— 任务集合 S 中的一个任务；

τ_s —— s 对应的权值。

6.3.4 每秒浮点运算次数

人工智能算力中心每秒浮点运算次数 C_{FP} 的定义见 3.8，见式 11。

注 1：在本文件中，本指标仅对可变负载训练测试有效。

注 2：在表述测试结果时标明位宽（如 FP16，FP32 等）。

$$C_{FP} = \sum_i \sum_j \frac{(N_F C_F + N_B C_B)}{T_{TR-ij}} \quad (11)$$

式中：

i —— 正整数，训练节点编号；

j —— 非负整数，模型结构变异次数（j=0,表示使用种子模型）；

C_F —— 特定模型上，前向传播过程计算量；

C_B —— 特定模型上，后向传播过程计算量；

N_F —— 特定模型和训练过程中，前向传播过程数量；

N_B —— 特定模型和训练过程中，后向传播过程数量；

T_{TR-ij} —— 节点 i，对模型 j 的训练用时。

6.3.5 每秒整型运算次数

人工智能算力中心每秒整型运算次数 C_{INT} 的定义参考 3.8，见式 12。

注 1：在本文件中，本指标仅对可变负载训练测试有效。

注 2：在表述测试结果时标明位宽（如 INT16，INT8，INT4 等）。

$$C_{INT} = \sum_i \sum_j \frac{(N_F C_F + N_B C_B)}{T_{TR-ij}} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (12)$$

式中：

i —— 正整数，训练节点编号；

j —— 非负整数，模型结构变异次数（j=0,表示使用种子模型）；

C_F —— 特定模型上，前向传播过程计算量；

C_B —— 特定模型上，后向传播过程计算量；

N_F —— 特定模型和训练过程中，前向传播过程数量；

N_B —— 特定模型和训练过程中，后向传播过程数量；

T_{TR-ij} —— 节点 i，对模型 j 的训练用时。

7 可靠性要求

7.1 基础要求

7.1.1 加速器可靠性

人工智能算力中心的人工智能加速处理器，满足以下可靠性要求：

a） 应支持以下故障的检测，并提供容错方案实现：

1） 加速器内部模块失效；

2） 加速器板供电模块失效，板掉电；

b） 应支持或通过使用 CPU 的功能支持模型保护；

c） 应支持 a）或 d）中提出故障造成错误或计算任务中断后的恢复执行；

d） 应支持多比特内存 ECC 故障检测；

e） 如加速器配备固件，宜支持操作系统带内、带外双升级通道；

f） 宜支持硬件唯一秘钥，并提供秘钥可靠存储机制；

g） 宜支持数据可靠使用，对用户选定数据的提供防篡改支持；

h） 宜支持关闭物理调试接口或其它防止非授权使用的机制；

i） 宜支持数据加解密和秘钥管理；

j） 宜支持至少一种国密算法；

k） 宜支持不可靠算法的标识，及禁止不可靠算法执行的机制；

l） 如具备物理资源的虚拟化能力，则宜实现任务级资源隔离；

m） 宜支持主芯片对可靠外设的身份认证；

n） 宜支持硬件漏洞 CVE 机制（如漏洞披露，描述，影响范围，涉及硬件型号和修复方法等）。

7.1.2 节点可靠性

人工智能算力中心的人工智能服务器节点，满足以下可靠性要求：

a） 应支持异常掉电，操作系统崩溃，磁盘、内存、中央处理器出错造成节点不可用后的计算任务恢复执行（恢复可由配套软件辅助完成，配套软件可包含操作系统等）；

b） 应支持扩展配备 RAID 0/1/10/5/50/6/60 冗余；

c） 风冷训练服务器应支持风扇模组热插拔，N+1 冗余；
d） 使用风冷散热时，风扇 COP 宜不小于 4.5；
e） 使用液冷散热时，宜支持水温不小于 45℃的液冷散热；
f） 配备硬件模块（如 BMC），支持系统参数监控、警告触发和远程管理；
g） 支持基于 UEFI 的故障管理；
h） 单节点宜支持带锁的服务器机箱面板；
i） 宜支持基于硬件可信根的可信启动；
j） 支持 ECC 1bit 纠错，ECC 2bit 报错；
k） 单训练服务器节点，应额外满足以下要求：
1） 支持非系统硬盘热插拔；
2） 支持 RAID 缓存（cache）；
3） 支持交流电源模块热插拔，支持 N+N 冗余；
4） 支持风冷或液冷散热，能在 5℃～35℃下工作；
5） 支持基于电容的系统掉电数据保护；
l） 单推理服务器节点，应额外满足以下要求：
1） 应支持交流电源模块热插拔，支持 1+1 冗余；
2） 应至少支持风冷，能在 5℃～35℃下工作；
3） 宜支持基板管理控制器模块使用业务网口（如基于边带管理（NCSI））。

7.1.3 网络可靠性

人工智能算力中心内的网络连接，应满足以下可靠性要求：
a） 在参数面，至少支持以下连接故障的检测，并提供容错方案：
1） 交换机（TOR/Leaf，EOR/Spine）不工作；
2） 线缆连接阻断；
3） 网卡不工作；
b） 在可实施的范围内，使用直连铜缆完成短距离传输；
c） 在可实施的范围内，使用光纤完成长距离传输；
d） 使用 Peer-Link 链路时，采取多条链路聚合的方式实施；
e） 使用 MLAG 成员接口时（MLAG 主设备连接用户侧主机或交换设备的 Eth-Truck 接口），采取 LACP 模式的链路聚合；
f） 使用堆叠组网方案时，由多台物理交换机虚拟成一台逻辑交换机，实现链路备份；
g） 支持 Overlay 层面的双活网关；
h） 支持基于 VRRP 协议的网关设备选举机制；
i） 支持汇聚交换机，核心交换机的冗余组网设计，单台交换机故障不影响作业执行。

7.1.4 算力中心整体及其它部件可靠性

人工智能算力中心整体及其它部件，满足以下可靠性要求，使用户应用无需对人工智能算力中心的可靠性实施加强措施：
a） 算力中心整体通过硬件、软件协作，支持以下可靠性特性要求：
1） 容错性控制模块在集群系统 Master 内设计，通过如 Kubernetes 高可用，YARN-RM 主备部署等，方案提供容错性；
2） 通道解耦，容错控制仅依赖所检视资源的状态，避免多控制源；
3） 故障报告通道，监视通道和控制通道解耦；

4) 支持训练现场的保存和恢复，实现无损训练；
5) 支持节点或通信不可用时，重调度新节点及配置集合通信，继续计算任务；
6) 支持集群模型恢复能力，能够寻找集群中模型相互冗余的节点，并在节点模型损坏时从其冗余节点上恢复模型；

b） 宜支持集群统一分布式缓存，能在故障时，不丢失数据；
c） 应支持设备状态（如启停、可用配置等）或错误码的查询；
d） 应支持人工智能加速器状态检查工具；
e） 应支持中心内部高速互联状态检查工具；
f） 应支持容错策略配置和实施；
g） 应具备故障分级，并实现基于故障级别的容错策略配置和实施；
h） 应支持故障后亲和性重调度；
i） 应支持节点故障检测；
j） 应支持面向容器的设备状态查询和报送；
k） 应支持针对特定故障（见表 2）的设备复位或修复；
l） 应支持针对特定故障（见表 2）的设备自动隔离；
m） 应支持集群训练任务的断点续训，能自动检测、隔离故障资源，在故障时保存断点信息（如checkpoint）并能调度冗余资源从故障断点恢复训练，全过程实现自动化；
n） 训练服务器机柜宜支持节点的水（如进水/出水）、电(如供电/信号)的盲插，避免因错误接入导致的故障。

7.2 扩展测试负载

人工智能算力中心可靠性的测试，应使用6.2.1规定的负载。在负载执行中，注入表2定义的故障，观察系统反应和恢复情况，测量、计算指标。

表 2 人工智能算力中心故障

模块	故障模式	故障原因	故障影响
D.人工智能加速器	1.片上内存多比特ECC	片上内存颗粒存储空间失效	对应单加速器不可用
	2.人工智能加速器故障	人工智能处理器芯片内部模块失效	对应单人工智能加速器不可用
	3.人工智能加速器板异常掉电	人工智能处理器板供电模块失效	整节点不可用
E.节点服务器硬件	1.宕机	异常掉电，操作系统崩溃，磁盘、内存、CPU错误	整节点不可用
F.网络设备	1.加速设备不可调用	人工智能加速器网卡故障，TOR/Leaf交换机故障，或其连接线缆断连	单人工智能加速器不可用，使用该加速器的IP作为检测IP时，报网络错误
	2.交换设备不可用	EOR/Spine交换机故障或其线缆断连	一个或多个加速器报网络错误

7.3 扩展指标

7.3.1 平均故障恢复时间

平均故障恢复时间（ms）是人工智能算力中心在执行特定任务时，中心的某部分或整体多次（≥3）发生同一故障而无法继续执行任务的时点T_{F1}，与该故障被修复，任务重新获得执行的时点T_{F2}，之间的差的平均值，见式13和式14。

注：修复方法一般包含自动修复和手动修复。

$$T_F = T_{F2} - T_{F1} \quad (13)$$

$$\overline{T_F} = \sum T_F \quad (14)$$

7.3.2 收敛比

收敛比（%）是人工智能算力中心中特定交换机层E接入的总实际带宽BW_{E-IN}与该层交换机总流出实际带宽BW_{E-OUT}的比值，见式15。

注：收敛比是可选指标。

$$C_E = \frac{BW_{E-IN}}{BW_{E-OUT}} \times 100\% \quad (15)$$

参 考 文 献

[1] Devlin, J., et. al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019(1): 4171-4186.

[2] Ren, Z., et. al. AIPerf: Automated machine learning as an AI-HPC benchmark[J]. Big Data Mining Anal, 2020 4(3): 208-220.

[3] Zeng, W. et. al. PanGu-alpha: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation[J]. arXiv, 2021 abs/2104.12369.

[4] Feurer, M. et. al. Efficient and Robust Automated Machine Learning[J]. Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015(2): 2755-2763.

[5] Giladi, R. and Ahituv, N.; SPEC as a Performance Evaluation Measure[J]. Computer, 1995, 28(8): 33-42.