团体标准

发 布

中国电机工程学会

20XX—XX—XX实施

20XX—XX—XX发布

微服务智能适配技术规范

Microservice intelligent adaptation technology specification

（送审稿）

T/CSEE XXXX—YYYY

代替 T/XXXX

ICS 19.020

CCS K85

目 次

[前 言 3](#_Toc17213)

[1 范围 4](#_Toc31681)

[2 规范性引用文件 4](#_Toc9058)

[3 术语和定义 4](#_Toc808)

[4 缩略语 5](#_Toc16712)

[5 架构要求 5](#_Toc2624)

[6 功能要求 6](#_Toc4677)

[6.1 接口定义 6](#_Toc19104)

[6.2 部署策略 6](#_Toc1215)

[6.3 适配组合 6](#_Toc28544)

[7 评价指标要求 7](#_Toc23509)

[7.1 性能评价指标 7](#_Toc26939)

[7.2 效能评价指标 8](#_Toc325)

[7.3 资源占用评价指标 8](#_Toc4647)

[8 测试方法 8](#_Toc18288)

[8.1 测试硬件及测试数据 8](#_Toc22261)

[8.2 测试分析 9](#_Toc9401)

[8.3 测试方法 9](#_Toc26411)

前 言

本文件按照《中国电机工程学会标准管理办法（暂行）》的要求，依据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。本文件的发布机构不承担识别专利的责任。

本文件由中国电机工程学会提出。

本文件由中国电机工程学会电力信息化标准专业委员会技术归口和解释。

本文件起草单位：北京中电普华信息技术有限公司、北京航空航天大学、国网信息通信产业集团有限公司、北京工业大学、北京国网信通埃森哲信息技术有限公司、国家电网有限公司、国家电网有限公司大数据中心、国网宁夏电力有限公司、北京电力交易中心有限公司、吉林电力交易中心有限公司、宁夏电力交易中心有限公司、国网山西信息通信分公司、中关村智联软件服务业质量创新联盟、北京嘀嘀无限科技发展有限公司、河南科评信息技术有限公司、山东中创软件商用中间件股份有限公司、浙江大学滨江研究院。

本文件主要起草人：于卓、郝庆利、吕海、吴文峻、梁堉、王思宁、崔蔚、黄晓光、范旭东、朱克、周春雷、李洋、王翰林、张显、韩松、乔宁、马军伟、张琪、任健、张奎、彭天豪、李隽、丁国富、崔忠峰、王潇、韩锋、潘晓华、李晓冬、赵恩恩。

本文件首次发布。

本文件在执行过程中的意见或建议反馈至中国电机工程学会标准执行办公室（地址：北京市西城区白广路二条1 号，100761，网址：http：//www.csee.org.cn，邮箱：[cseebz@csee.org.cn](mailto:cseebz@csee.org.cn)）。

微服务智能适配技术规范

1. 范围

本文件规定了电力行业人工智能领域微服务智能适配的架构要求、功能要求、评价指标要求、测试方法。

本文件适用于电力行业人工智能领域微服务智能适配建设，也可为电力行业人工智能领域微服务产品的研发、运维和使用提供参考。

1. 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　5271.1-2000　信息技术　词汇　第1部分：基本术语

GB/T　5271.31-2006　信息技术　词汇　第31部分：人工智能　机器学习

GB/T　5271.34-2006　信息技术　词汇　第34部分：人工智能　神经网络

Q/GDW 12119-2021 微服务架构设计导则

T/IQA 12-2021 智能微服务适配框架

T/IQA 13-2021 智能微服务适配效能评价指标体系

1. 术语和定义

GB/T 5271.1-2000、GB/T 5271.31-2006、GB/T 5271.34-2006、Q/GDW 12119-2021和T/CES 128-2022中界定的以及下列术语和定义适用于本文件。

人工智能 artificial intelligence

一门交叉学科，通常视为计算机科学的分支，研究表现出与人类智能（如推理和学习）相关的各种功能的模型和系统。

微服务 microservices

一些可协同工作、小而自治、分布式模式的软件服务，以服务方式实现的软件包，具有部署独立、通信轻量的特点，支撑单一业务逻辑的功能实现。

微服务智能适配 microservice intelligent adaptation

将人工智能模型和算法引入微服务的全生命周期中，通过智能化方法，采用一定的策略对服务进行集成和组合，实现微服务系统的开发、部署、运行和维护；还可以通过微服务架构及工具，对智能模型进行管理，实现智能模型的微服务化部署和应用。

容器 container

将单个微服务的资源划分到孤立的组中，以便更好的在孤立的组之间平衡有冲突的资源使用需求。

服务组合 service portfolio

定义为服务策略和服务设计服务组合涵盖了一个组织中所有服务所有信息，每一服务都表明其状态和历史。

任务 task

被调度的训练/推理对象。

节点 node

智能微服务平台中的工作节点，运行真正的智能微服务程序。

机器学习模型 machine　learning　model

采用机器学习方法建立的输入与目标输出联系的计算模型。

机器学习引擎 machine　learning　engine

提供机器学习开发及运行的计算组件。

1. 缩略语

下列缩略语适用于本文件。

CPU：中央处理器（Central　Processing　Unit）

GPU：图形处理器（Graphic　Processing　Unit）

AI：人工智能（Artificial　Intelligence）

CICD：持续集成（Continuous　Integration）和持续部署（Continuous　Deployment）

I/O：输入/输出（Input/Output）

2. 架构要求

电力行业人工智能领域微服务智能适配的架构见图1，包括对微服务智能适配过程中的功能要求、评价指标要求和测试方法进行标准化。可以服务于电力行业人工智能领域微服务产品和应用的研发、运维和使用。

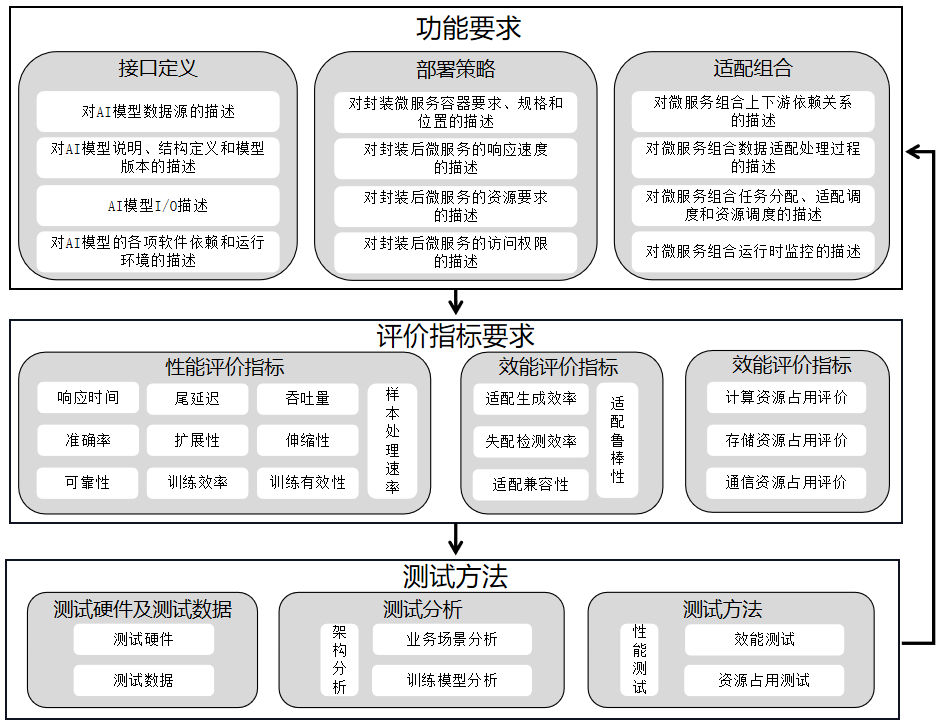


图1 架构要求

电力行业人工智能领域微服务智能适配的架构要求为：

1. 应支持将来自不同AI模型库，具有不同编译和运行环境的AI模型，封装成为微服务；
2. 应支持在管理大区或互联网大区的微服务环境中，集成组合异构微服务智能适配模型，形成智

能化的服务流程，同时实现CICD；

1. 应支持电力行业主流的AI算法，主流的机器学习模型以及主流的机器学习引擎，并支持根据实际的电力业务需要和用户的实际习惯，提供分析、开发、交互接口和开发文档支持等功能。同时各模块间的接口应遵循电力行业常见的架构和协议。
2. 通过对微服务智能适配进行解构分析，根据QoS属性特征，针对微服务之间耦合互联、动态匹配的质量，评估分为微服务的性能评价指标、微服务智能适配评价指标、资源占用评价指标。
3. 功能要求
   1. 接口定义

接口定义是指将AI模型和算法利用容器技术封装成微服务，并规范微服务组件的各接口定义，具体包括：

1. 对AI模型数据源的描述；应支持来自信息业务系统或通过离线数据采集的数据源；
2. 对AI模型说明、结构定义和模型版本的描述；应给出模型的具体规则；
3. 对AI模型I/O的描述，说明当前模型的输入输出信息；应支持根据具体电力业务需要定义AI

模型的I/O属性；

1. 对AI模型的各项软件依赖和运行环境的描述；应支持不同需求下的对AI模型软件依赖和运

行环境的定义；应支持在不影响现有微服务正常运行的前提下，部署新版本的微服务组件；应

支持微服务组件的增删启停、版本管理、服务历史记录、服务当前状态的查询等。

* 1. 部署策略

部署策略是指为多个微服务智能组件在组合部署过程中的策略配置及对应的优化方法，具体包括：

1. 对封装微服务容器要求、规格和位置的描述；应支持用户根据不同的功能需求，配置容器的标准化要求（如容器运行时标准及容器镜像标准）、规格（如服务的生命周期等）和部署位置；
2. 对封装后微服务的响应速度的描述；应支持用户根据不同的功能需求，配置微服务组件的响应速度；
3. 对封装后微服务的资源要求的描述；应支持用户根据不同的功能需求，配置微服务组件的资源要求（如CPU、GPU、内存需求等）；
4. 对封装后微服务的访问权限的描述；应支持用户根据不同的功能需求，配置微服务组件的访问权限，控制其所能调用的微服务。
   1. 适配组合

适配组合是指将组合好的微服务整合成完整的规范的智能业务流程，服务于电力行业，实现更好的管理维护，具体包括：

1. 对微服务组合上下游依赖关系的描述；应支持不同微服务组合在不同电力业务要求下，根据其不同的上下游依赖关系将对应节点形成不同的流程组合；
2. 对微服务组合数据适配处理过程的描述；应提供数据适配的功能，根据不同的输入数据，将节点自动适配，形成对应的微服务组合；
3. 对微服务组合任务分配、适配调度和资源调度的描述；应支持对微服务任务分配和流程调度的功能，支持针对不同电力应用场景下的应用；
4. 对微服务组合运行时监控的描述；应提供运行时监控的功能，包括日志管理，可以根据日志进行故障定位及排查；及提供针对监控指标及日志的告警能力，同时对主要监控指标的可视化展示功能。
5. 评价指标要求
   1. 性能评价指标
      1. 在线推理

性能评价要求如下：

1. 响应时间
   1. 响应时间应为接收到服务请求时刻到系统输出推理结果时刻的时间间隔；
   2. 在系统负载小于 80%时，推理调用响应时间不超过 3 秒；
   3. 在系统负载小于 80%时，前端页面响应时间不超过 5 秒；
   4. 在系统负载小于 80%时，后端服务响应时间不超过 3 秒；
   5. 在系统负载小于 80%时，算法训练任务创建、推理服务申请应时间不超过3分钟。
2. 尾延迟
   1. 尾延迟是满足某业务需求所需最后一个微服务的响应时间；
   2. 针对满足该业务所需所有服务度量响应时间，取最大值；
   3. 在不同领域的应用场景下，尾延迟均值及99分位值应低于该领域定制阈值。
3. 吞吐量
   1. 满足业务要求的情况下，单位时间内整个系统的处理能力；
   2. 吞吐量越大表示微服务系统在线推理处理能力越强。
4. 准确率
   1. 预测正确的业务量占总业务量的比率；
   2. 准确率越高表示微服务能提供更准确的推理能力。
5. 扩展性
   1. 满足一定时间内业务发展的需求的程度；
   2. 满足硬件、软件、系统扩展升级过程不影响现有样本数据、训练任务、模型推理服务；
   3. 扩展性越高表示微服务系统能在未来一段时间内提供更好的服务能力。
6. 伸缩性
   1. 在流量高峰期支撑对外服务的能力；
   2. 伸缩性越强标识微服务能更好应对流量瞬时激增的场景。
7. 可靠性
   1. 在不能抗力环境下，应满足7×24小时服务不中断，提供冗余的网络设备、通信线路、系统硬件，保证容错率和可用性；
   2. 可靠性越高表示微服务系统能在未来一段时间内提供持续的服务能力。
      1. 离线训练/测试

性能评价要求如下：

1. 样本处理速率
   1. 单位时间处理的样本数；
   2. 在保证训练效能，要求处理样本的速率越高越好。
2. 训练效率
   1. 训练特定迭代次数的用时；
   2. 为保证训练效率，要求系统具有快速的迭代。
3. 训练有效性
   1. 训练达到特定精度的用时；
   2. 为保证模型有效性，要求训练结果达到高精度。
   3. 效能评价指标
4. 适配生成效率
   1. 评价微服务智能适配业务流程的成功率及花费时间；
   2. 自动生成业务流程成功率越接近100%越好，花费时间越少越好。
5. 失配检测效率
   1. 评价自动识别失配业务流程的数量占所有失效业务的比率，及其失效时刻到发现时刻的时间；
   2. 失配业务自动识别的比例越接近100%越好，自动发现的越快越好。
6. 适配兼容性
   1. 评价当运行环境发生变化，或服务版本发生演化时，已有业务流程的兼容情况；
   2. 当运行环境发生变化时，不兼容情况出现是的业务失败的发生次数占比越接近0越好。
7. 适配鲁棒性
   1. 评价在一定时间周期内，当输入发生错误的变化，使得已有业务流程未按预期完成，出现失效的数量；
   2. 单位时间周期内，出现失效的密度越低，适配的鲁棒性越好。
   3. 资源占用评价指标
8. 计算资源占用评价

微服务智能适配使用的计算单元（如：CPU和GPU）完成计算任务的时间，即计算资源占用时间。

1. 存储资源占用评价

微服务智能适配使用的存储单元（如内存和存储）在一定时间内完成的数据读写的数据量。

1. 通信资源占用评价
   1. 微服务智能适配通信在特定时间点或时间段内的通信数据量的大小及占用整体网络带宽的大小；
   2. 对资源占用的要求是在给定的业务目标下，资源占用时间越少越好，资源的占用分布越均衡越好。
2. 测试方法
   1. 测试硬件及测试数据
      1. 测试硬件

测试环境包括系统接口定义、部署策略、网络环境、适配场景，具体测试环境应基于至少500个服务容器实例，准备至少包括50个服务注册实例的服务环境；运行微服务系统、训练任务、推理服务等组件的服务器应具备GPU算力资源，支持CUDA和CUDNN加速。

* + 1. 测试数据

针对微服务智能适配应用场景，测试数据应覆盖微服务智能适配的核心电力业务场景，并从智能适配覆盖、智能业务覆盖以及功能性能角度进行设计，选择至少50种的微服务构建测试集合。

* 1. 测试分析
     1. 架构分析

1. 分析微服务智能适配整体部署架构，包括通信网关部署、负载均衡部署、容器平台部署、容灾处理机制、扩容方案等；
2. 分析微服务智能适配使用的技术架构。
   * 1. 业务场景分析
3. 分析微服务智能适配应用的电力业务场景，针对具体业务场景制定测试方案；
4. 分析业务场景交易链路，旨在验证微服务智能适配在服务链路上的适配情况。
   * 1. 训练模型分析
5. 分析微服务智能适配训练模型，针对训练模型制定测试方案；
6. 在微服务智能适配训练模型预期内制定相应的测试场景以验证训练模型有效性。
   1. 测试方法
      1. 性能测试

性能测试方法如下：

1. 针对响应时间、尾延迟、吞吐量、样本处理速率和训练效率，依据本标准上述定义在服务运行

时对相应指标进行实时监控、记录和汇总计算；

1. 针对训练有效性，依据上述定义在微服务执行过程中记录推理结果精度及训练时间，达到相应精度后停止训练，记录训练时间；
2. 针对结果准确率和训练有效性，依据上述定义在微服务执行过程中记录推理结果，业务结束后进行推理结果准确性判定及相关计算；
3. 针对扩展性，依据上述定义的微服务性能评价指标，模拟不同业务量对微服务进行可用性测试，达到特定指标后结束，记录相关指标；
4. 针对伸缩性，依据上述定义的微服务性能评价指标，模拟流量波峰期以及波谷期对服务进行可用性测试，达到特定指标后结束，记录相关指标。
5. 针对可靠性，依据上述定义的微服务性能评价指标，模拟网络设备、通信线路、系统硬件对服务进行可靠性测试，达到特定指标后结束，记录相关指标。
   * 1. 效能测试
6. 适配生成效率：在给定的多个应用场景下，记录给定的业务适配生成的数量、成功数量及所花费的时间，并依据上述定义计算相应生成效率；
7. 失配检测效率：在注入服务适配的故障后，记录导致失效适配的业务流程的总数量、被自动发现的数量及执行时间，并依据上述定义计算相应的失配检测效率；
8. 适配兼容性：在切换给定数量的智能业务运行环境或服务版本后，记录发生服务失效的数量，并计算相应的失效密度；
9. 适配鲁棒性：在使用给定数量的不准确的输入信息后，记录导致业务流程失效的数量，并计算相应的失效密度。
   * 1. 资源占用测试

通过监控软件和设备，在服务适配及业务运行的过程中记录所使用计算资源、存储资源及通信资源 的占用时间。

1. 计算资源占用测试

要求在指定配置的服务器上进行测试，需要提前指定服务器CPU、内存、内核等相关配置，固定变量，选取正常运行期、流量波峰期、流量波谷期至少三种场景进行压力测试，统计记录计算资源消耗情况。

1. 存储资源占用测试

要求基于真实业务场景，预估线上最大流量，以线上最大流量作为单位流量持续发起压力测试，达 到在较短时间验证微服务在长期运行过程中对存储的占用情况。

1. 通信资源占用测试

要求固定带宽、网卡等网络资源配置，基于真实业务场景，选取正常运行期、流量波峰期、流量波 谷期至少三种场景进行压力测试，统计记录不同场景下通信资源占用情况。