

基于 Ray 框架的 AI 工业检测 SaaS 平台研究

Research on AI Industrial Inspection SaaS Platform Based on Ray Framework

陈亚峰,徐玉龙,张黎明,刘江锋(中讯邮电咨询设计院有限公司,北京 100048)

Chen Yafeng, Xu Yulong, Zhang Liming, Liu Jiangfeng (China Information Technology Designing & Consulting Institute Co., Ltd., Beijing 100048, China)

摘要:

介绍了 AI 工业检测 SaaS 平台的设计背景及设计目标,提出了基于 AI 图像检测原子能力和检测流程编排进行图像集中化处理的检测模式;详细阐述了 AI 工业检测平台的工作流程;分析了平台实现方案中涉及的图像文件共享访问、大规模异构 GPU 集群下资源调度等技术问题,并提供了使用 Ray 开源框架解决上述问题的实现方案。

关键词:

Ray; AI; 工业检测; 图像处理

doi: 10.12045/j.issn.1007-3043.2025.08.014

文章编号: 1007-3043(2025)08-0069-06

中图分类号: TN919

文献标识码: A

开放科学(资源服务)标识码(OSID):



Abstract:

The design background and objectives of the AI industrial inspection SaaS platform are introduced, a detection mode based on AI image detection atomic capabilities and detection process orchestration for centralized image processing are proposed. The workflow of the AI industrial inspection platform is detailed. An analysis of technical issues such as image file sharing access and resource scheduling under large-scale heterogeneous GPU clusters in the platform implementation scheme is provided, along with an implementation solution using the Ray open-source framework to address the aforementioned issues.

Keywords:

Ray; AI; Industrial inspection; Image processing

引用格式: 陈亚峰,徐玉龙,张黎明,等. 基于 Ray 框架的 AI 工业检测 SaaS 平台研究[J]. 邮电设计技术, 2025(8): 69-74.

1 背景

计算机视觉技术在工业上应用领域广阔,基于计算机视觉技术的工业检测相关应用产品已在工业领域应用多年。核心功能包括测量、检测、识别、定位等。在工业检测领域,与人类视觉相比,机器视觉具有精度高、速度快、稳定性高等优势^[1-2];目前基于机器视觉的工业检测产品已经得到了广泛的应用。

目前国内最常用的工业检测产品主要基于 Halcon 及 VersionPro^[3]等商用机器视觉产品进行二次开

发,在多数情况下能够满足客户需求,但随着 AI 深度学习技术^[4]及 5G 工业互联网等通信能力的发展,原有的工业检测产品也逐渐暴露出一些不足^[5]。

a) 现有工业检测产品主要针对单机运行模式研发^[6],主要应用于生产线单机场景,不适合大规模、平台化的运营模式;无法实现大型企业内部或不同企业间的工业检测资源和能力的共建共享。

b) 现有工业检测产品主要基于传统计算机视觉技术,对近年来迅猛发展的深度学习技术支持不足。

c) 相对于开源的 OpenCV 机器视觉框架及开源的深度学习技术而言,现有工业检测产品价格昂贵,成本较高,从而限制其应用范围。

收稿日期: 2025-06-30

d) 缺乏检测结果的大数据分析能力。单机模式导致检测数据的格式无法统一和规范,与其他系统(如MES、ERP等系统)集成度较低,无法整合分析数据。

对于大型生产企业和大型工业园区等组织而言,更希望能够运营平台化的工业检测平台,以集中利用软硬件资源,特别是在GPU等硬件资源成本日益昂贵的情况下;而平台化的运营模式,也有助于企业降低总体的工业检测实施成本。

2 平台目标及需求

基于AI能力的工业检测SaaS平台的主要研发目标如下。

a) 以SaaS云服务的方式向用户提供工业检测能力。用户在生产线本地只需部署图像采集设备、控制设备,无需再部署工业检测用计算机及单机模式的工业检测软件。

b) 平台可提供多种传统机器视觉检测能力和基于AI的工业检测能力,并且可以进行能力的扩展,以形成可扩展的检测能力仓库。

c) 平台具备对工业检测能力的编排功能。用户可通过对多种检测能力灵活编排,快速满足工业检测需求。

d) 具备集中的检测数据存储及数据分析能力。

3 平台架构设计及业务流程

3.1 平台架构设计

平台功能模块^[7]及架构如图1所示。

平台基础设施层包括GPU服务器、CPU服务器及TPU推理服务器等。平台主要采用Kubernetes作为基础资源管理平台,使用KubeRay开源工具在Kubernetes平台中部署基于Ray框架研发的各类应用组件,同时将Docker技术作为部分业务承载环境的补充。平台主要构建了2类基础检测能力仓库:一类是深度学习视觉能力仓库,这类检测能力主要使用深度学习技术完成传统机器视觉难以完成的检测任务,一般会用到GPU或TPU资源;另一类是传统机器视觉能力仓库,主要是基于OpenCV/Halcon等技术完成简单的图像预处理任务,如缩放、反转、锐化等,此类任务使用CPU即可完成。通过检测流程编排器的方式将检测基础能力进行组合、协同及调用,从而完成特定的工业检测任务。SaaS用户服务主要面向用户及租户提供业务管理配置、统计分析查看等业务功能。图像输入及控制输出适配管理主要通过插件化的方式支持主流工业相机通信协议及工业总线协议,便于快速与现场工装环境对接。

由于采用Kubernetes作为基础设施环境,平台能够根据业务容量需求进行硬件资源的水平扩展;一般

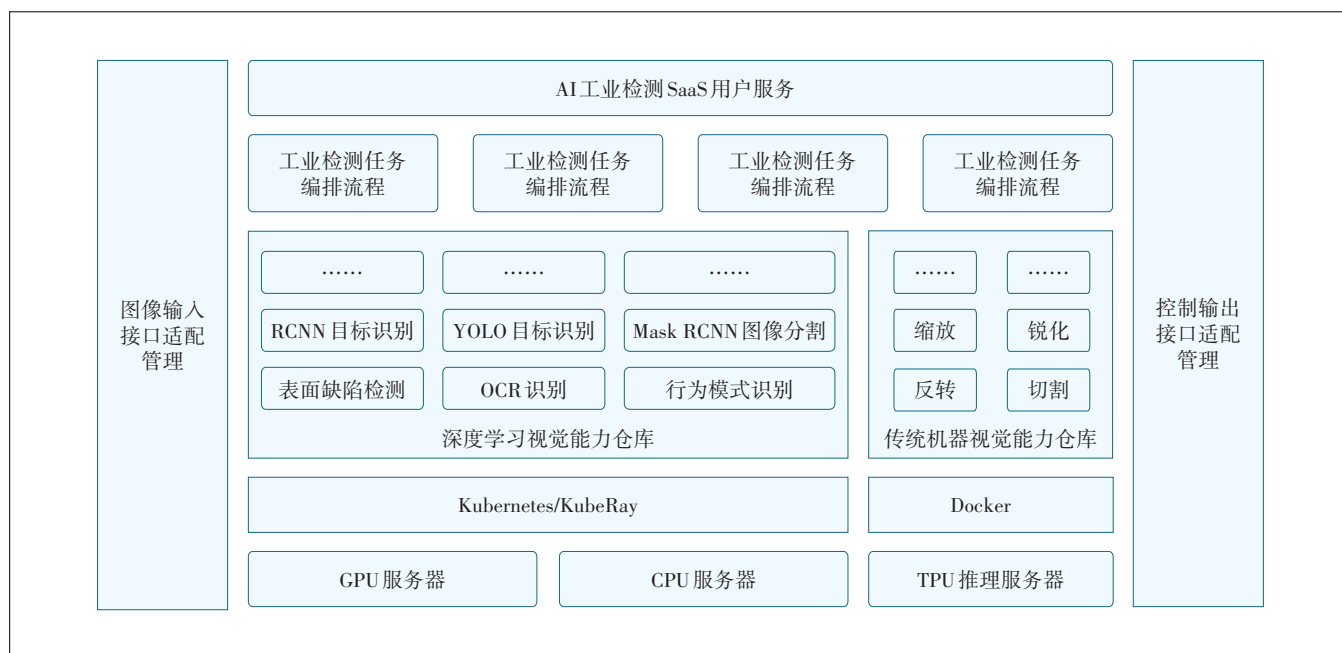


图1 平台功能模块及架构

要求配置3台以上CPU服务器、1台以上的GPU服务器作为集群基础设施平台。

3.2 典型检测业务处理流程

基于AI能力的工业检测平台部署方案与传统单机工业检测系统不同,在车间现场仅配备有检测设备、工业相机、指示灯、报警器、开关等产线终端设备,不具备图像处理用计算资源及相应的检测软件;产线终端设备通过5G或有线网络与检测中心连接;检测平台部署在企业或园区的数据中心内。基于上述部署方案的核心检测业务流程如图2所示。

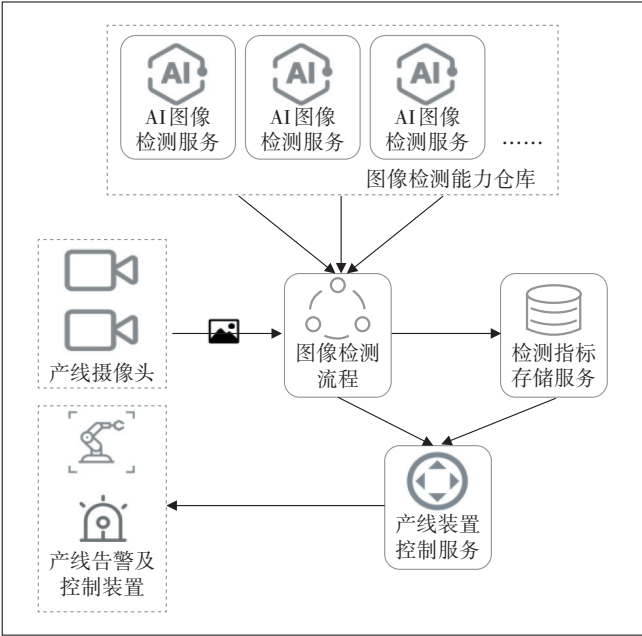


图2 AI工业检测SaaS平台业务流程

- a) 部署在产线上的摄像头获取图像。
- b) 将获取到的图像通过数据通信网络上传至AI工业检测平台。
- c) 平台提供可扩充的AI图像检测原子能力仓库,可通过编排AI检测原子能力形成图像检测流程。
- d) 产线图像被输入图像检测流程运行实例,并按照编排流程对图像进行检测,检测结果存入集中的检测指标存储服务器,同时通知产线装置控制服务检测结果。
- e) 远程控制告警及控制装置向产线反馈检测结果。
- f) 产线终端根据检测结果进行通知、告警或控制动作。
- g) 保存检测结果数据,用于生产经营分析及用户

业务应用。

3.3 平台业务适应性分析

AI工业检测SaaS平台采用平台化、集中化的图像检测处理方式,主要业务处理环节耗时及相应技术要求如表1所示。

a)平台采用集中处理图像的模式,图像需通过网络传输。因此,在满足检测需求的情况下产线采集图像参数越低越好。一般情况下,图像分辨率不高于1 024×768^[4],且应尽量采用灰度图像采集方式^[8]。在大部分情况下,单个图像大小不宜超过200 KB;在低速产线条件下,可放宽图像采集尺寸、色彩等方面的要求。

b) 平台可采用多种网络接入方式,根据实际产线检测任务的需求,可考虑采用5G专网^[9]、Wi-Fi、千兆有线等多种方式连接平台与产线设备,其延迟等特性如表1所示。

表1 AI工业检测SaaS平台处理环节详表

序号	处理环节	耗时/s	备注
1	图像采集	<0.03	一般工业相机帧率不低于30 FPS
2	图像传输	5G专网	≈0.03 下行500 Mbit/s,上行100 Mbit/s,文件200 KB
		Wi-Fi	≈0.01 上下行800 Mbit/s,抖动较大、延迟较高、文件200 KB
		千兆有线局域网	≈0.01 上下行1 000 Mbit/s,抖动小、延迟低、文件200 KB
3	图像处理	机器视觉图像处理原子服务	<0.01 不使用神经网络及GPU方式处理图像
		AI深度学习图像处理原子服务	0.01~0.05 使用神经网络及GPU方式处理图像
		图像处理流程编排及调度	<0.03 处理流程调度、图像处理结果逻辑判定等
4	控制反馈	<0.01	向产线反馈图像处理结果

c) 平台通过对原子图像处理能力进行编排来完成具体的检测任务,检测任务可以复用原子图像处理能力。单个任务的复杂程度由检测业务需求涉及的检测项目数量决定;单个检测项目一般使用3~5个图像处理原子能力即可完成;如果任务存在多个检测项目,可根据需要并行处理。

基于上述分析及在某空调外观检测产线上的实际案例,AI工业检测SaaS平台较适合应用在中低速产线业务场景^[10],其单次检测任务时延要求在100 ms以上;对于高速产线(如每秒处理图像超过30帧),则宜采用本地检测处理方式。

4 基于 Ray 框架的技术实现方案

Ray^[11]是一个基于内存共享的分布式计算框架,由 UC Berkeley 的 AMPLab 团队创建。Ray 是首个实现训练、仿真和服务统一的分布式计算框架。Ray 框架目前主要用于 AI 大规模深度学习的训练领域,例如 OpenAI 的 ChatGPT 大语言模型就是基于 Ray 框架完成训练的。根据对 AI 工业检测 SaaS 平台技术实现需求的分析,Ray 框架的特性^[12]较为适合应用于该平台的研发。

4.1 检测图像的存储及访问

产线摄像头获取图像后,需将图像传送至数据中心以完成检测任务。AI 检测原子服务(如图像识别、图像分割、图像锐化等)部署在数据中心内不同的服务器或容器中,获取的图像需在这些 AI 检测原子服务间频繁传递及访问;此外,部分图像处理服务将会对原始图像进行处理并生成新的图像,这些新的图像同样需要作为其他图像处理服务的输入。通常的解决方案如图 3 所示。

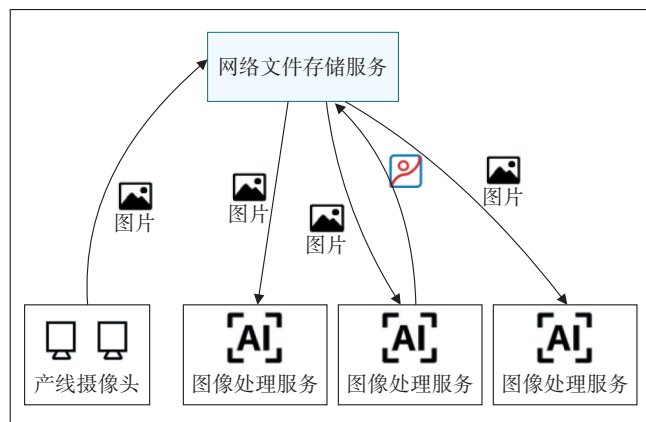


图3 基于网络文件存储服务的图像文件访问方案

采用网络文件存储服务(如 NFS、CIFS 等)保存产线摄像头产生的图片,部署在计算集群中的 AI 图像处理服务通过网络访问图像并执行检测、处理等任务。采用该方式存在的主要问题包括:

- 通过网络文件系统拷贝图像文件,效率较低。
- 各个图像处理服务读取图像均需通过文件系统访问,同一个图像文件需要在不同的图像处理服务间通过网络多次传递。
- 需通过复杂的图像缓存编程逻辑,才能实现部署在同一服务器上的图像处理服务共享图像文件。

使用 Ray 开源框架的分布式对象存储机制^[13]来处理图像在 AI 图像处理服务间的传输和访问,可较好地解决上述问题。Ray 开源框架的分布式对象存储技术方案如图 4 所示。

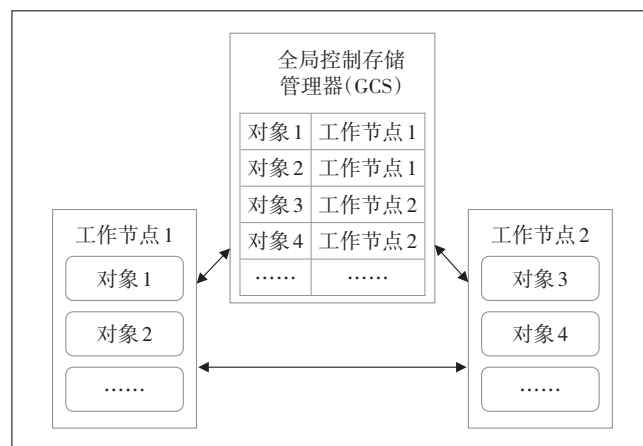


图4 Ray 框架的分布式对象存储技术方案

Ray 使用全局控制存储管理器(GCS)保存整个集群中的对象存储信息,并通过内存共享机制在每个节点上实现了一个对象存储器(Object Store),从而使在同一个节点运行的任务之间不需要拷贝就可以共享数据。当一个任务的输入不在本地时,则会在执行之前将它的输入复制到本地的对象存储器中,且仅需复制一次。

使用 Ray 框架的分布式对象存储来保存工业检测平台中的图像数据,能够解决采用网络共享文件存储方案带来的各种问题。AI 图像检测服务程序可以使用 Ray 框架提供的 API 接口访问图像数据,而无需关心图像数据在集群中存储及复制的细节^[14],从而降低了研发难度。使用 Ray 框架分布式对象存储与网络文件系统(CIFS)完成同样工作任务的性能对比如图 5 所示。

4.2 AI 图像检测服务的调度运行

工业检测平台的检测能力仓库可容纳并扩展多种图像检测能力服务。不同的检测能力对硬件资源的需求不同。基于深度学习的图像检测能力通常需要昂贵的 GPU 资源,而基于 OpenCV 等传统图像处理软件的检测能力仅需 CPU 就能运行。为了提高检测平台集群的硬件利用率、降低运行成本,需要平台支持根据检测能力的资源需求,将检测能力调度到特定的集群工作节点中运行。Ray 框架对该研发需求提供了较好的支持。

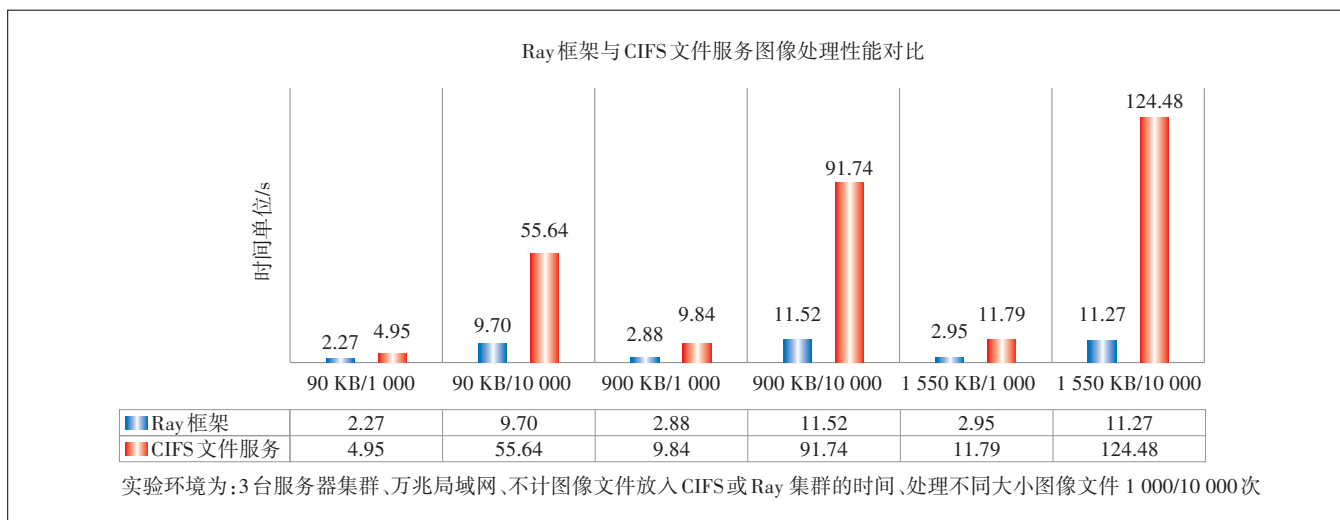


图5 Ray 框架与 CIFS 共享文件服务图像处理性能对比

4.2.1 放置组(Placement Groups)方案

为了提高硬件资源利用率, Ray 框架提供了放置组(Placement Groups)方案^[12]来支持计算任务的灵活编排和调度。 Ray 框架放置组资源分配示意如图 6 所示。

a) Ray 集群中的工作节点在初始化时,可对节点的 CPU、GPU、内存等资源信息进行配置。

b) 可以在集群中创建具备不同资源规格的放置组(Placement Groups)。需注意的是,放置组分配的资源以列表形式呈现,即同一规格的放置组在集群中可以创建多个实例形成资源池。

c) 在工业检测平台中,单个图像检测能力在实现上对应 Ray 开源框架中的 Actor^[15]。在 Actor 启动运行时,可以指定其使用的放置组, Ray 框架将会调度该

Actor 在合适的资源放置组上运行。其以 Python 代码实现的主要逻辑示例如图 7 所示。

使用 Ray 放置组方案,可以较好地解决工业检测平台中的图像检测能力的调度运行问题,实现集群资源利用的最大化。

4.2.2 异构 GPU 资源调度方案

在工业检测平台服务器集群中,不同厂家、不同型号的异构 GPU 资源(如英伟达 GPU 与华为 GPU 混用)并存是常见情形。仅通过前述放置组方案中的 GPU 数量信息无法区分不同型号、厂家的 GPU 类型,不能满足平台根据 GPU 类型精细化调度图像处理任务运行的需求。

在 Ray 框架中,提供了自定义资源标签功能。可以在工作节点初始化时,自定义资源标签信息,用来

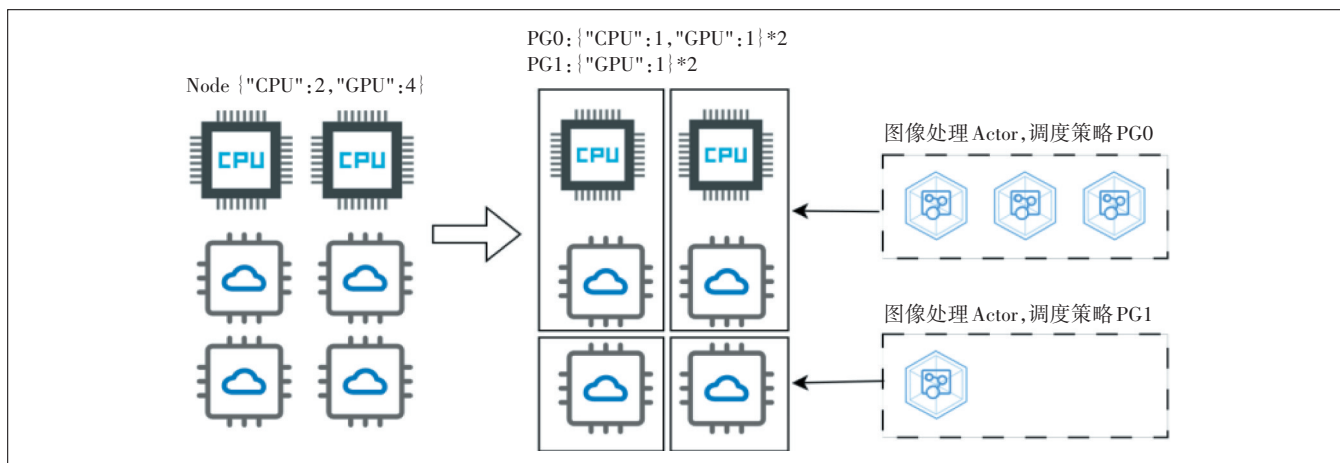


图6 Ray 框架放置组资源分配示意

```
# 导入 ray 开源包
import ray
# 初始化工作节点,2CPU,4GPU
ray.init(num_cpus=2,num_gpus=4)
# 创建放置组
pg0=placement_group([{"CPU":1,"GPU":1}])
pg1=placement_group([{"GPU":1}])
.....
# 图像处理服务 Actor 实现样例
@ray.remote(num_cpus=1,num_gpus=1)
class ImgProcessActor:
    def __init__(self):
        pass
    def ready(self):
        pass
# 创建图像处理服务 Actor 并通过 Ray 框架调度至合适的放置组运行
actor=ImgProcessActor.options(
    scheduling_strategy=PlacementGroupSchedulingStrategy(
        placement_group=pg0,
    )
).remote()
.....
```

图7 Python代码实现的主要逻辑示例

扩充标记工作节点具备的资源特性(见图8)。

```
# 初始化工作节点,2CPU,4个昇腾910 GPU
ray.init(num_cpus=2,num_gpus=4,resources={"ascend910":4})
#创建放置组
pg0 = placement_group([{"CPU":1,"GPU":1,"ascend910":1}])
```

图8 自定义资源标签

上述方案借助自定义资源标签,能让 Ray 放置组识别华为“昇腾910”型号的GPU资源。

Ray 框架通过自定义资源标签方案增强了资源调度度的灵活性,使工业检测平台能够根据 GPU 型号、GPU 显存大小等因素来调度 AI 图像检测服务能力的运行,较好地解决了异构 GPU 资源的调度问题。

4.3 平台实现成果

充分利用开源 Ray 框架的相关特性,以较低的研发成本解决了 AI 工业检测 SaaS 平台研发中涉及的技术难点,最终实现了平台研发的目标。

5 结语

工业检测 SaaS 平台采用了平台化运营模式,即工业园区或企业的所有工业检测需求在一个平台化系统中完成。这种方式解决了单机运营模式下的种种缺陷,更能适应 5G 等高速网络、大规模云计算环境、确定性网络、GPU 集群等新技术环境下的工业检测需求。平台采用“AI 检测能力仓库”,不但可以解决当下企业面临的工业检测需求,还可使企业在工业检测能

力方面不断积累和提升。

通过充分利用开源 Ray 框架的技术特性,该平台解决了大规模异构 GPU 集群下的资源调度和利用、图像数据在集群内部的高效传输等问题。同时,平台采用“检测能力仓库+检测流程编排器”的集中化工业检测模式,充分复用了软硬件资产,对业界同类平台的实现具有一定的参考意义。

参考文献:

- [1] 彭炜轩. 基于计算机视觉的工业检测技术研究[J]. 现代制造技术与装备, 2023(7):112-114.
- [2] 陈吉成,许热,王大伟. 计算机视觉技术下的工业检测技术分析[J]. 软件, 2023(12):154-156.
- [3] 陈洋,张林. 工业视觉检测系统及应用[J]. 科技展望, 2015(16):119-121.
- [4] 王国超,柴琳,金立左. 目标检测算法[J]. 工业控制计算机, 2023(12):10-14.
- [5] 陈秋霞. 机器视觉技术在工业检测领域中的应用[J]. 设备管理与维修, 2018(16):140-142.
- [6] 樊慧超. 机器视觉技术在工业检测中的应用[J]. 数字通信世界 2020(12):156-157.
- [7] 王翔. 高并发访问下的分布式系统架构设计[J]. 智能城市, 2023(3):5-7.
- [8] 杨航,陶青川. 红外图像目标检测算法 YOLO-TC[J]. 现代计算机, 2023(9):56-59.
- [9] 吴锋,张继超. 基于 5G 的智能工业检测高性能云服务系统[J]. 信息记录材料, 2023(4):86-88.
- [10] 张杰. 基于改进 Faster-RCNN 的小目标检测[J]. 现代计算机, 2023(14):14-18.
- [11] Ray Official WebSite[EB/OL]. [2025-02-20]. <https://ray.io>.
- [12] Ray-Project[EB/OL]. [2025-02-20]. <https://github.com/ray-project/ray>.
- [13] 赵康,马陈燕,王道军. 基于 Ray 并行分布式框架的深度强化学习计算平台[J]. 软件, 2022(11):179-183.
- [14] 王雅娴,奚文娟. 一种用于分布式系统的多级缓存策略[J]. 现代计算机, 2023(19):12-16.
- [15] 侯伟. 基于 Ray 框架的分布式计算研究[J]. 现代信息科技, 2023(23):65-68.

作者简介:

陈亚峰,高级工程师,学士,主要从事数据通信网络工程咨询设计及 NFV、SDN 相关的研发工作;徐玉龙,工程师,硕士,主要从事软件研发、数字孪生及元宇宙方向的产品规划、解决方案设计及项目研发交付工作;张黎明,工程师,学士,主要从事智慧城市、工业互联网方向的产品规划、解决方案设计及研发工作;刘江锋,工程师,学士,主要从事工业互联网方向的产品规划、解决方案设计及项目实施交付工作。