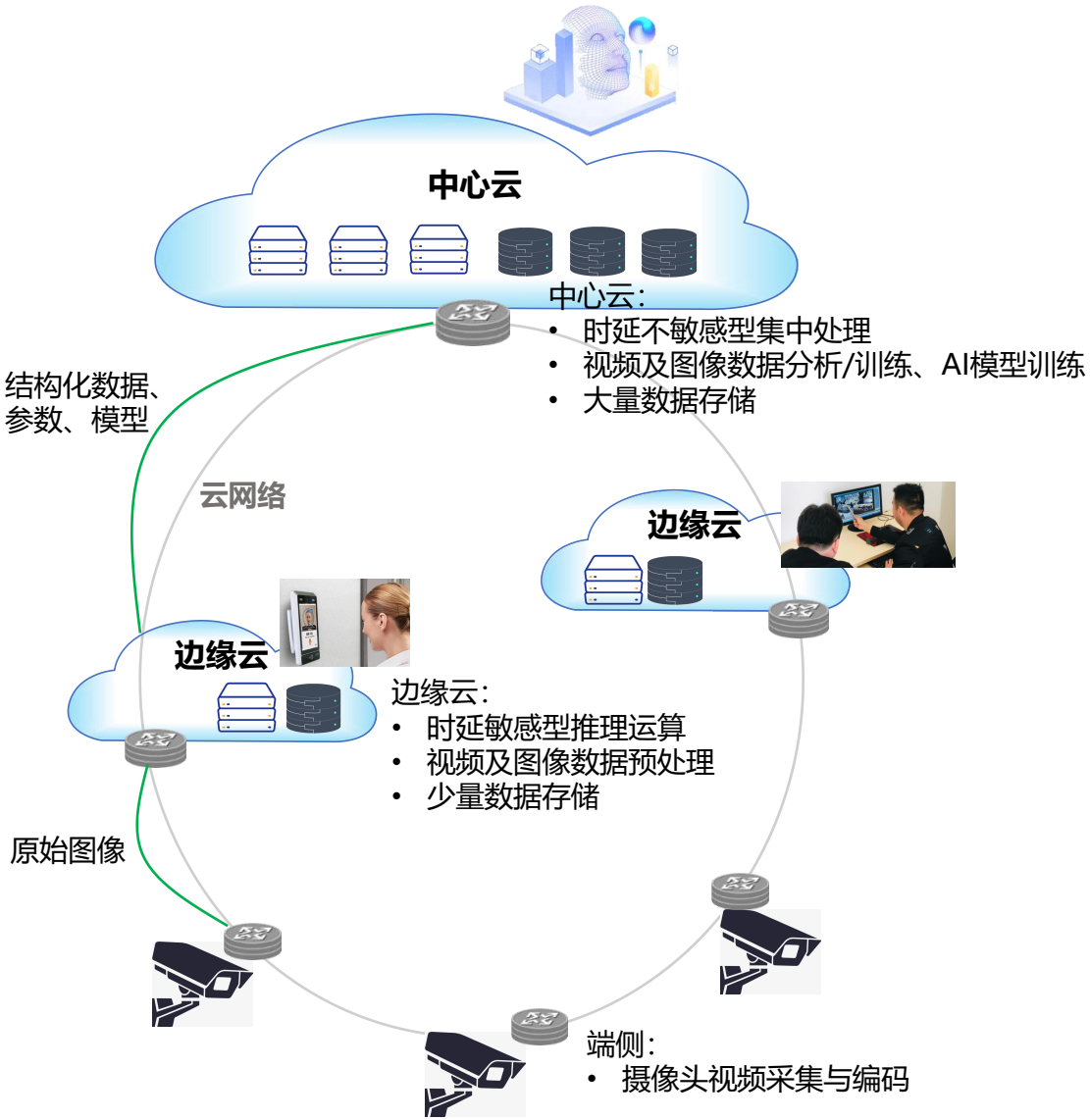
算力网络-云边协同典型业务场景——视频/图像处理



场景描述：

- 基于算力网络构建视频/图像监控采集+推理+处理系统，处理实时性强、数据量大的视频及图像业务
- 云、边、端侧相互协同，端侧收集数据、边缘侧进行实时推理运算及数据预处理、中心侧负责数据分析/模型训练及数据存储
- 根据客户需求，在边缘（客户近端）动态选择推理模型部署位置，在端侧与边缘、边缘与中心动态度调整网络路由及带宽

适合业务：

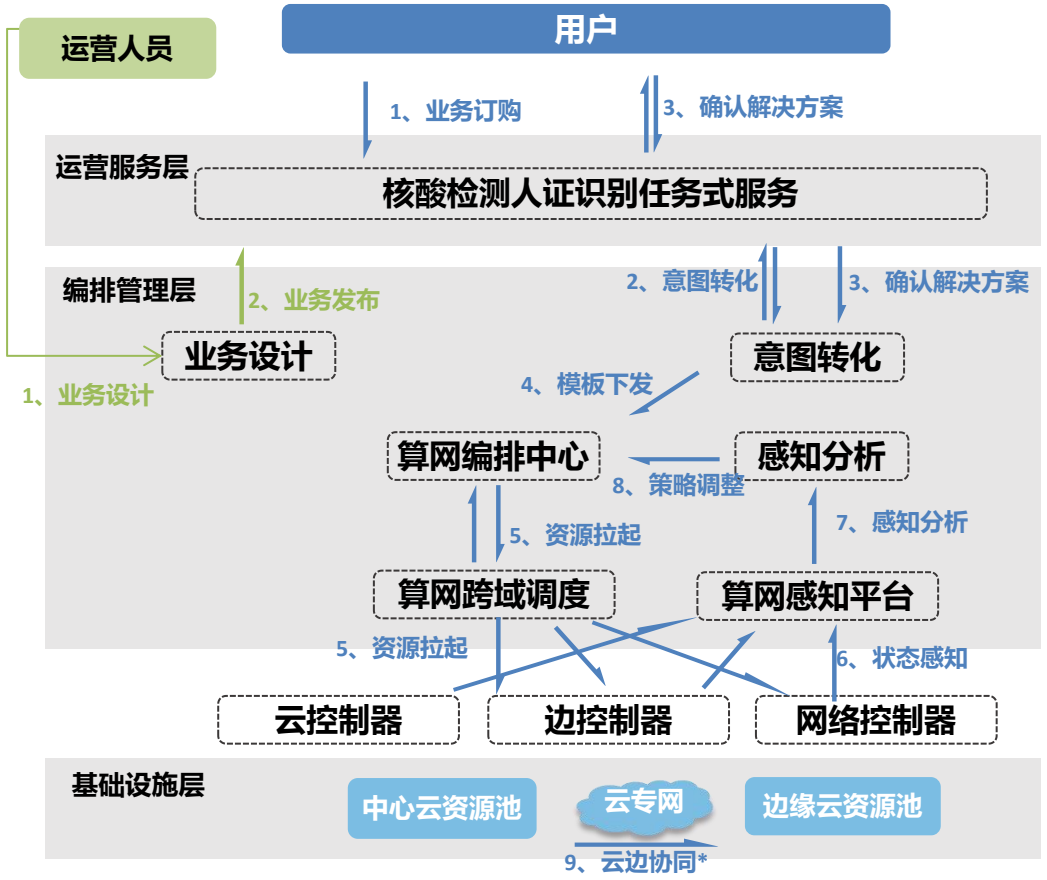
- AI训练、远程驾驶、自动驾驶、视频监控&检测、智能工控等
- 可统一归类为**“中训边推”类业务**

算力网络需求：

- **边缘侧引入GPU**等高性能算力，推理运算业务支持**跨架构部署**
- **中心侧引入GPU等高性能算力**完成数据处理及模型训练
- 低时延、高带框网络，支持**网络带宽灵活调整、路由灵活选择**

端到端业务流程示意——中训边推

中训边推指由中心云进行大规模模型训练，在边缘进行数据近场或现场推理的业务，如人证识别、视频检测、智能工控、车联网等。算力网络通过分布式资源统一管理、云边协同调度、对业务和资源的智能感知分析和对应用的敏捷部署，支持中训边推类业务的快速应用。



人证识别任务式服务示例（中训边推流程）

■ **设计态**（**绿色流程**）

- **1. 业务设计**：面向运营/产品设计人员，通过托拉拽或模板化的方式快速构建**人证识别任务式服务**模型
- **2. 业务发布**：业务设计构建的“人证识别服务”在算力网络运营平台解决方案专区上架，接收用户订购，同时算网服务随任务启动而订购开通、随任务完成而退订回收，并支持分钟、流量等更灵活的计费策略

■ **运行态**（**蓝色流程**）

- **1. 业务订购**：用户向运营平台提出业务需求，包括图片大小、检测数量、倾向的检测地点、成本约束等参数
- **2. 意图转化**：意图转化实现业务需求到资源需求的转译，根据图像处理的任务规模、实时性要求、检测地点、成本约束等条件匹配资源方案，并生成多种解决方案，传递给运营平台
- **3. 用户确认**：运营平台根据营销策略生成最终报价单，用户确认最终解决方案，运营平台将用户确认的方案专递给算网大脑
- **4. 模板下发**：将解决方案模板（json或yaml格式）下发到算网编排中心，算网编排中心完成模板解析及ABCDNETS多要素能力调用
- **5. 资源拉起**：通过算网跨域调度，对接云控制器、边控制器和网控制器API，完成资源开通、训练/推理服务部署、任务启动等一体化交付
- **6. 状态感知**：持续采集图像识别业务状态和底层资源状态，为感知分析提供数据支撑
- **7. 感知分析**：将业务和资源数据实时发送给感知分析，感知分析根据业务和资源实时状态综合考虑资源位置、成本、业务利用率、网络性能，生成动态跨域的算网扩缩容策略
- **8. 策略调整**：感知分析将调整策略发送至算网编排中心，充分利用云原生业务的灵活敏捷特性，进行业务的负载均衡、应用的快速拉起、算网资源的更新等，确保人证识别任务目标达成
- **9. 云边协同***：在中心云训练的模型持续优化迭代，并实时将最新训练模型推送至边缘

中训边推业务流程及算网需求映射分析

需求1：（开发态，针对业务开发者）
- AI业务开发设计支持跨架构运行与迁移（至少包含CPU、GPU，ASIC/FPGA可选）
- 编排管理层支持业务镜像、配置文件等内容上传/存储

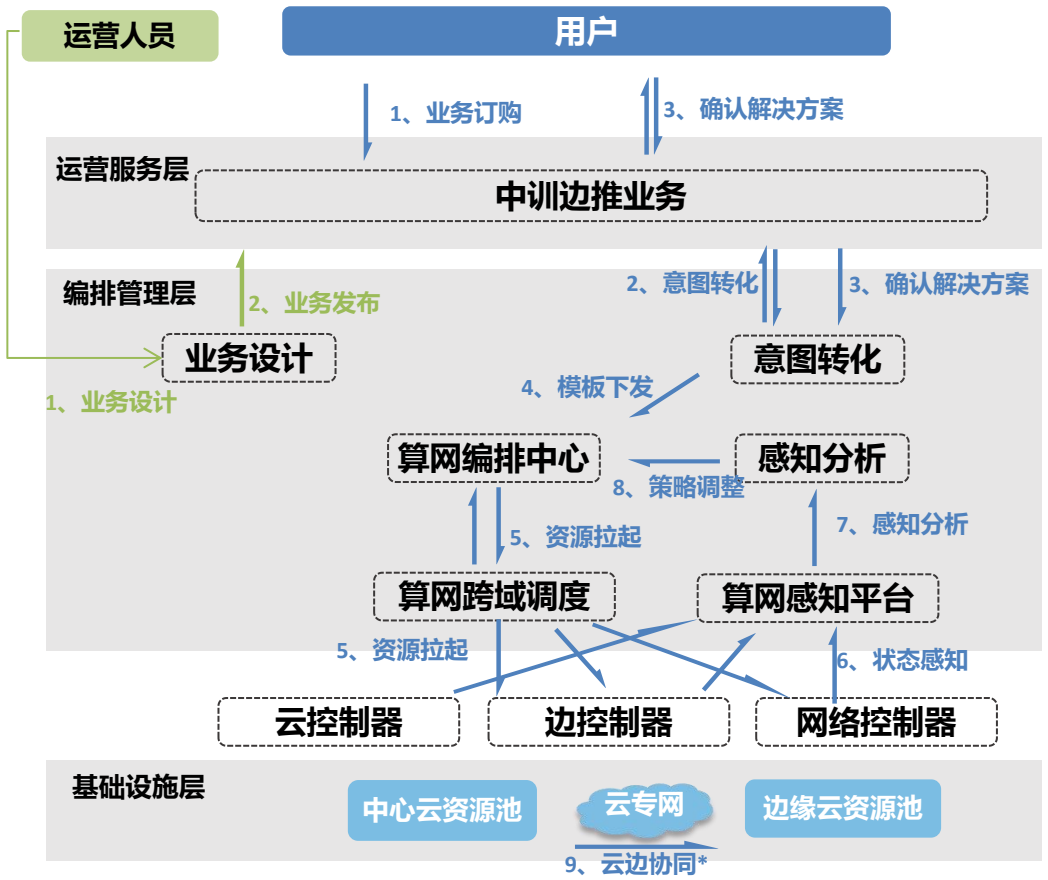
需求2：（业务设计、业务发布）
- 编排管理层提供蓝图功能，支持灵活业务设计
- 编排管理层提供业务生命周期控制相关接口
- 运营服务层支持业务服务模式配置，并与编排管理层业务生命周期控制功能协同

需求3：（业务订购）
- 运营服务层提供用户Portal，支持用户选择业务，并填写业务需求

需求4：（意图转化）
- 业务开发者提供业务处理单元任务所需要的性能及算力资源参考，以及任务量增加时算力资源需求计算模型
- 编排管理层纳管算网资源并实时监控资源状态
- 编排管理层根据用户的业务需求+业务单位任务的资源需求计算总资源需求，并结合算网资源状态确定解决方案
- 算力网络支持实现算力度量

需求5：（意图转化、用户确认）
- 运营服务层及编排管理层（下文简称两层）间有消息传递接口，供传递业务需求及解决方案

需求6：（模板下发）
- 编排管理层提供编排部署模板，描述各类算网资源需求



需求7：（资源拉起）
- 编排管理层协同云/边控制器，选择合适的中心云节点和边缘云节点（基于监控到的实时资源状态及网络状态）
- 编排管理层向各云节点推送业务部署文件，完成业务部署，并实现跨架构调度
- 编排管理层协同网络控制器完成网络配置

需求8：（状态感知）
- 编排管理层监控业务运行状态及业务所用资源运行状态

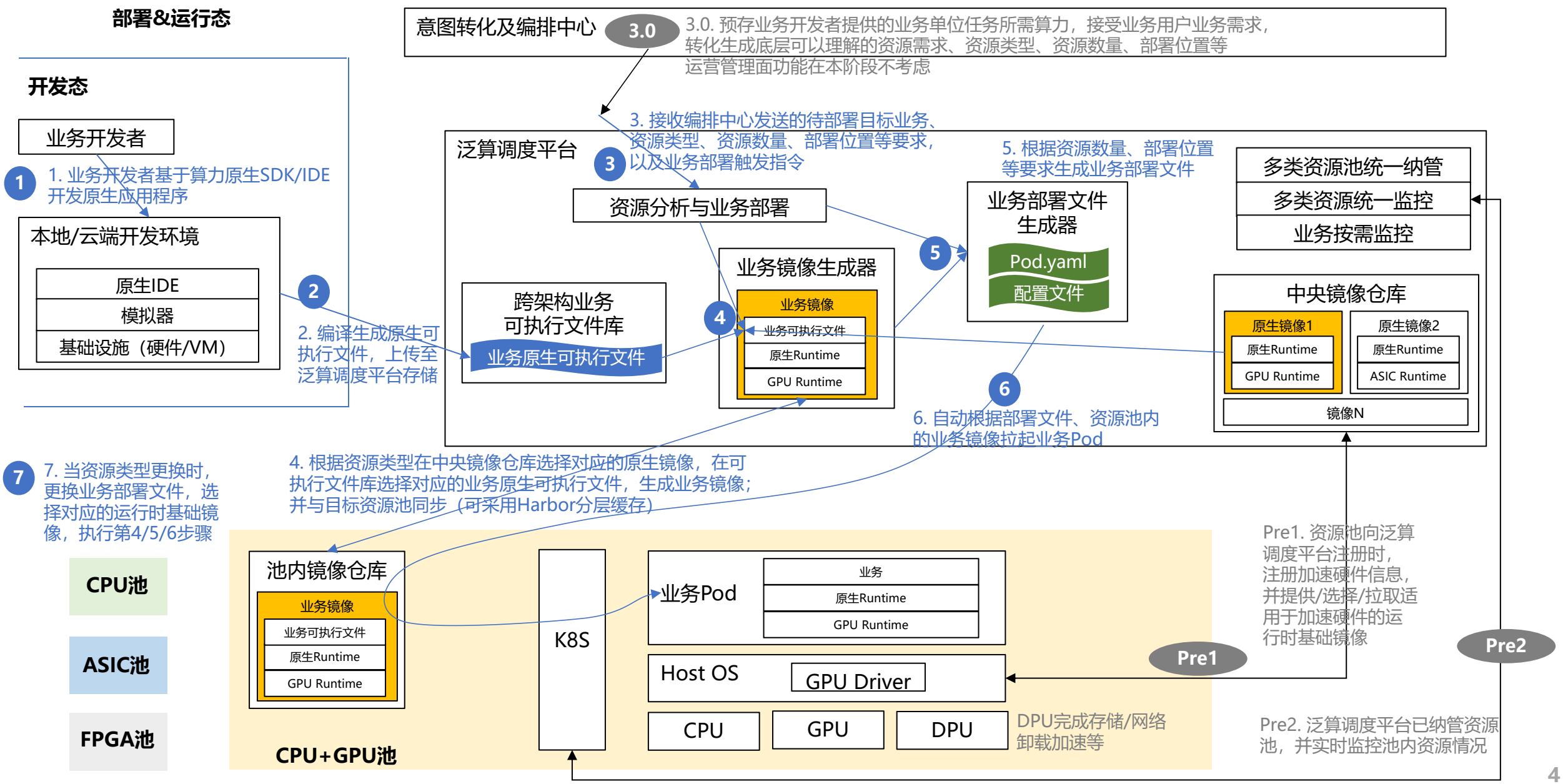
需求9：（感知分析）
- 业务提供基于业务运行状态及业务所用资源运行状态的动态控制策略
- 编排管理层基于业务动态控制策略、用户对业务的需求生成业务调整策略

需求10：（策略调整）
- 编排管理层基于业务调整策略自动调整业务

需求11：
基础设施层云资源池提供异构计算资源（至少包含CPU、GPU，ASCI/FPGA可选），以及存储资源

中训边推业务中最基础的需求是业务跨架构开发、部署、运行，第一阶段可基于该需求完成端到端设计与验证（增加：算力形态多样、跨架构调度的需求、准确调度到目标节点、网络转发性能需要提升算力处理能力）

基础端到端流程


部署&运行态
开发态
业务开发者
1
1. 业务开发者基于算力原生SDK/IDE开发原生应用程序
本地/云端开发环境
原生IDE
模拟器
基础设施（硬件/VM）
2
2. 编译生成原生可执行文件，上传至泛算调度平台存储
意图转化及编排中心
3.0
3.0. 预存业务开发者提供的业务单位任务所需算力，接受业务用户业务需求，转化生成底层可以理解的资源需求、资源类型、资源数量、部署位置等运营管理面功能在本阶段不考虑
泛算调度平台
3
3. 接收编排中心发送的待部署目标业务、资源类型、资源数量、部署位置等要求，以及业务部署触发指令
资源分析与业务部署
跨架构业务可执行文件库
业务原生可执行文件
4
业务镜像生成器
业务镜像
业务可执行文件
原生Runtime
GPU Runtime
5
5. 根据资源数量、部署位置等要求生成业务部署文件
业务部署文件生成器
Pod.yaml
配置文件
多类资源池统一纳管
多类资源统一监控
业务按需监控
中央镜像仓库
原生镜像1
原生Runtime
GPU Runtime
原生镜像2
原生Runtime
ASIC Runtime
镜像N
6
6. 自动根据部署文件、资源池内的业务镜像拉起业务Pod
7
7. 当资源类型更换时，更换业务部署文件，选择对应的运行时基础镜像，执行第4/5/6步骤
4. 根据资源类型在中央镜像仓库选择对应的原生镜像，在可执行文件库选择对应的业务原生可执行文件，生成业务镜像；并与目标资源池同步（可采用Harbor分层缓存）
CPU池
ASIC池
FPGA池
池内镜像仓库
业务镜像
业务可执行文件
原生Runtime
GPU Runtime
K8S
业务Pod
业务
原生Runtime
GPU Runtime
Host OS
GPU Driver
CPU
GPU
DPU
DPU完成存储/网络卸载加速等
CPU+GPU池
Pre1
Pre1. 资源池向泛算调度平台注册时，注册加速硬件信息，并提供/选择/拉取适用于加速硬件的运行时基础镜像
Pre2
Pre2. 泛算调度平台已纳管资源池，并实时监控池内资源情况