AI 中台建设中智算资源的设计方案

林善亮，徐飞
北京中网华通设计咨询有限公司，北京市首都公路发展集团有限公司，北京 100073
DOI: 10.61369/SSSD.2025080039

摘　　要： 随着人工智能技术的迅猛发展，AI 中台作为企业智能化转型的核心基础设施[1]，其重要性日益凸显。在 AI 中台的架构中，智算资源是支撑模型训练、推理服务及大规模数据处理的关键要素。如何科学、高效地设计智算资源，直接关系到 AI 中台的性能、成本与可持续发展能力[2]。本文阐述了 AI 中台建设中智算资源的设计方案，涵盖设计原则、需求分析、资源计算过程、部署策略等关键环节。

关 键 词： 设计原则；需求分析；计算过程；资源部署

Design Scheme of Intelligent Computing Resources in AI Middle Platform Construction

Lin Shanliang, Xu Fei
Beijing Zhongwang Huatong Design Consulting Co., Ltd., Beijing Capital Highway Development Group Co., Ltd., Beijing 100073

Abstract: With the rapid development of artificial intelligence technology, the AI Middle Platform, as the core infrastructure for enterprises' intelligent transformation[1], has become increasingly prominent in its importance. In the architecture of the AI Middle Platform, intelligent computing resources are a key element supporting model training, inference services, and large-scale data processing. How to design intelligent computing resources scientifically and efficiently is directly related to the performance, cost, and sustainable development capability of the AI Middle Platform[2]. This paper expounds on the design scheme of intelligent computing resources in the construction of the AI Middle Platform, covering key links such as design principles, demand analysis, resource calculation process, and deployment strategy.

Keywords: design principles; demand analysis; calculation process; resource deployment

前言

智算资源是 AI 中台的“心脏”与“引擎”，为模型训练、推理预测、数据预处理等计算密集型任务提供强大的算力支持[3]。与传统的通用计算资源不同，智算资源通常以 GPU、TPU、NPU 等专用加速器为核心，具备高并行计算能力、大内存带宽和低延迟通信特性，能够高效处理深度学习等复杂 AI 算法[4]。

在 AI 中台的架构中，智算资源主要承担以下职能：

模型训练：支持大规模数据集上的深度神经网络训练，通常需要高算力、高内存和高速网络[5]。

模型推理：提供低延迟、高吞吐的在线服务，满足实时性要求高的业务场景。

数据处理：加速数据清洗、特征工程、向量化等预处理任务，提升数据准备效率[6]。

资源调度：通过虚拟化、容器化等技术，实现算力资源的动态分配与共享，提高资源利用率。

一、设计原则

在仅进行微调训练的场景下，算力规划需聚焦“轻量化部署、弹性伸缩、国产化适配与成本效能平衡” 四大核心原则[7]，通过以下策略实现资源最优配置。

（一）轻量化算力架构

训练层：采用中等算力硬件（如24GB-32GB 显存级设备），通过参数高效微调技术（如 LoRA）优化模型顶层参数，仅需训练 0.1% 参数量即可注入行业知识，单卡即可完成百亿级模型48小时内微调，硬件投入降低90%。

推理层：微调后立即进行模型压缩（INT8量化 + 知识蒸馏），将模型体积缩减至1/4，使边缘计算设备（如15W 低功耗国产芯片）可承载千路视频实时分析，时延严格控制在200ms 内。

（二）弹性资源调度

基于容器化编排平台实现动态扩缩容：在线推理服务根据并发请求量（QPS）自动启停实例，应对突发流量（如节假日客服

请求激增3倍）[8]；

建立“潮汐算力复用”机制：日间资源优先保障高并发推理任务，夜间空闲算力自动切换至批量微调任务（如政策适配训练）[9]，资源利用率从40%提升至75%。

（三）国产化与安全协同

硬件自主可控：训练与推理层国产化芯片承载比例＞60%，非敏感任务（如车流预测）由国产高性能芯片处理，敏感任务（含隐私数据）在加密芯片分区运行；

零信任防护：微调数据通过联邦学习实现本地化处理，推理服务间实施双向认证与动态授权，阻断横向渗透风险[10]。

二、需求分析

业务需求评估：明确AI应用场景（如智能交通、数据分析、预测建模等），确定算力需求（训练/推理比例、并发量、延迟要求等）[11]。

技术需求评估：根据AI框架（TensorFlow、PyTorch等）、模型规模（参数量、数据量）及算法复杂度，测算所需算力（CPU/GPU/TPU资源）[12]。

资源规划：制定短期（1年内）和长期（3-5年）算力扩展计划，预留弹性扩容空间。

三、智算资源计算过程

（一）多模态大模型训练智算资源计算

以多模态Qwen2.5-VL-72B为例，按照30天训练时长计算，多模态础模型的参数量是72B，峰值同时并发开10个训练任务，训练算力需求估算，分析如下：

（1）单个训练任务的总算力消耗(E)

以多模态该公式用于计算训练一个模型所需的总浮点运算次数（FLOPS）。

$E=6\times P\times T=6\times 72\times 10^{9}\times 2\times 10^{10}\approx 8.64\times 10^{21}$ FLOPs

E：总算力消耗，单位为FLOPs。

P：$P=72\times 10^{9}$。

T：训练过程中的总token数量，$T=2\times 10^{10}$。

其中E公式中的6，包括训练过程中前向传播2和后向传播4。

总算力8.64×10^{21} FLOPs换算成PFLOPs/day：

$E/(10^{15}\times 24\times 60\times 60)=100$ PFLOPs/day

所需算力卡数量(N)

根据以上得出每天要处理的算力资源，再除以单卡的算力，得出所需的GPU卡的数量。

以华为Atlas 800I A2 910B3显卡举例。

$N=E/(D\times F\_GPU\times U)=(8.64\times 10^{21})/(2592000\times 0.313\times 10^{15}\times 40\%)\approx 27$(卡)

N：所需的GPU卡数量。

E：单个任务的总算力消耗。

D：训练时长，单位为秒。在本例中，D=30天$=30\times 24\times 60\times 60=2592000$秒。

F_GPU：单张GPU的理论峰值算力，单位为FLOPS。以华为910B3为基准，其FP16 Tensor Core峰值算力约为313 TFLOPS，即0.313×10^{15}FLOPS。

U：硬件利用率。实际训练中，由于数据读取、通信开销等，GPU无法达到100%的峰值利用率。我们通常取一个保守的估计值，例如U=40%。

（2）训练10个并发任务显卡数量

由于需要同时运行10个相同的训练任务，总需求是单个任务的10倍。

总显卡数量：27张/任务×10个任务=270张华为910B3显卡。

（3）总结

在训练任务时长维度下，所需显卡数量与训练时长成反比。对于Qwen2.5-VL-72B模型，如果将训练时长缩短至15天，则需要54张显卡；若延长至60天，则只需要13.5张显卡。

在并行训练维度下，所需的总算力消耗和显卡数量与并行任务数量成正比。例如，若需同时运行10个Qwen2.5-VL-72B的训练任务，总算力消耗将达到8.64×10^{22}FLOPs，需要270张华为910B3显卡；若运行20个任务，则总算力消耗和显卡数量将分别增至1.73×10^{23}FLOPs和540张华为910B3显卡。

（二）多模态大模型推理智算资源计算

以多模态Qwen2.5-VL-72B为例，多模态大模型的推理一般和客户业务应用部署在一起。一般先推算占用多少显存，通过所需要多少总的显存，得出多少显卡。部署一个实例，显存占用公式如下：

GB显存 = 1.2*N*B参数

（1.2考虑开启K-V Catch缓存加速以及数据的batch批量输入）

N：模型FP16/BF16浮点精度，FP16（2Byte/参数），N=2；

B：为多模态大模型参数，取72；

GB显存 = 1.2*2*72=172.8GB

以华为Atlas 800I A2 910B3为例，单卡NPU910B3为64G显存。部署一个实例所需显卡：172.8/64=2.7（卡），向上取整为3张显卡。

如果部署5个实例，至少需要15张华为NPU910B3显卡。

（4）计算算力需求[11]

根据以上的参数，可分别计算保守估计和激进估计所需要的算力需求：

保守估计 =2*参数*350*15=$2*671*10^{9}*350*15=7.0455*10^{15}$（FLOPS）=7.0455 PFLOPS

激进估计 =2*参数*1050*15=$2*671*10^{9}*1050*15=21.137*10^{15}$（FLOPS）=21.137 PFLOPS

（5）结论：

折算成华为NPU910B3（313 TFLOPS = 0.313 PFLOPS/

卡）作为所需直观 GPU 数

低估：7.0455 / 0.313 = 23 张（华为 NPU910B3）

高估：21.137 / 0.313 =68 张（华为 NPU910B3）

四、智算资源部署

（一）部署模式选择

混合云架构：核心业务采用私有云（保障数据安全），弹性需求结合公有云（如突发训练任务）。

本地化部署：若数据敏感性高，采用自建数据中心或国产化算力集群（如昇腾、寒武纪芯片等）。

（二）硬件选型

训练层：高性能训练服务器（国产芯片），支持分布式训练。

推理层：低功耗 GPU/ASIC 芯片（国产芯片），优化实时推理效率。

存储与网络：配套高速存储和低延迟网络。

五、总结

本文系统阐述了 AI 中台建设中智算资源的设计方案，从设计原则、需求分析、资源计算过程到部署实施，构建了一套完整的理论方法。

首先，智算资源设计应遵循“轻量化部署、弹性伸缩、国产化适配与成本效能平衡”等核心原则[13]，确保平台的灵活性、稳定性与可持续性。其次，通过深入的业务场景需求、技术需求等，避免资源不足或过度配置[14]。在资源计算过程中，结合训练与推理的特性，建立数学模型估算 GPU 等硬件需求[15]。最后，在部署阶段，根据企业实际情况选择私有云、公有云或混合云架构。

参考文献

[1] 李航 . 人工智能：一种现代的方法 [M]. 北京：机械工业出版社，2020.

[2] 阿里云 . 混合云 AI 中台解决方案白皮书 [R]. 2022.

[3] 孙宪超 . 促进智算资源互联互通建立智算基础设施评价体系 [N]. 证券时报 ,2024-03-04(A05).

[4] 刘景磊，陈佳媛，李莹，等 . 中国移动 NICC 新型智算中心核心技术布局和展望 [J]. 通信世界 ,2023,(22):43-45.

[5] 孙长秋，杜长斌，李欣宇，等 . 智算中心关键技术研究 [J]. 通信管理与技 ,2024,(02):33-37+52.

[6] 郭慧 . 面向智算中心的多维资源智能协同调度关键技术研究 [D]. 北京邮电大学 ,2024.

[7] 史锋，陈骆颖，窦建华 . 兼容并存 " 合 " 而共舞——以 " 中台 " 为核心的浦东新区电子政府平台建设模式分析 [J]. 上海信息化 ,2009,(12):68-70

[8] 刘勤 . 数字电视节目中台标字幕叠加系统的建设 [J]. 视听界（广播电视技术）,2010,(03):47-50.

[9] 车品觉 . 建设数据中台，赋能创新改革 [J]. 新经济导刊 ,2018,(10):22-24.

[10] 郭全中 . 智媒体构建中的中台建设 [J]. 新闻与写作 ,2019,(11):71-75.

[11] 陈新宇 ." 中台 " 成为构架企业数字营销的主要模式 [J]. 中外管理 ,2019,(12):132-133.

[12] 数据中台为智慧城轨建设赋能 [J]. 城市轨道交通 ,2019,(12):56-57.

[13] 谭宇翔，顾盛楠 . 基于南水北调工程业务中台的微服务架构的设计与实施 [J]. 信息系统工程 ,2019,(10):38-39.

[14] 秦成 . 制造业企业中台建设思考与实践 [J]. 智能制造 ,2019,(Z1):50-53.

[15] 田鹏 . 数据中台在通信行业服务类企业中的应用研究 [J]. 中国新通信 ,2025,27(08):7-9.

