YRT SERVER WHITEPAPER 2026

AI GPU 服务器选型白皮书 2026

从真实工作负载出发，判断 GPU 计算服务器、AI 训练服务器与项目制 AI 平台的配置边界。

先判断任务

推理、微调、训练、多模态、科研和渲染对应不同平台。

再判断系统

显存、互联、数据路径、网络和机房条件一起决定可用性。

最后落到交付

正式方案要覆盖配置、测试、运维、质保和验收说明。

客户留资版 v0.2 · 更新日期：2026-06-04 · 英睿特 YRT Server

00. 这份白皮书解决什么问题

AI GPU 服务器采购最容易出现两个极端：只按 GPU 数量、单卡算力或最低价格做比较；或者直接套用别人项目的配置，没有结合自己的模型、数据、机房和交付要求。

这份白皮书不是固定报价单，而是一套采购前判断框架。您可以用它快速确认自己的需求更接近 G 系列 GPU 计算服务器，还是 T 系列 AI 训练服务器；也可以提前梳理预算、显存、互联、数据路径、机房和运维交付要求。

> **核心原则：先明确工作负载，再确定服务器类型；先确认交付环境，再落到具体配置。**

01. 一页判断：先看任务，不先看 GPU

典型任务	优先产品方向	采购时最该确认
本地大模型推理、RAG、知识库问答、API 服务	G 系列 GPU 计算服务器	显存、并发、模型数量、本地 NVMe、网络和长期运行稳定性
LLM 微调、视觉模型、多模态、AI 科研平台	T 系列 AI 训练服务器	高显存、GPU 互联、NCCL、训练软件栈、数据路径和满载验证
多课题组共享、学院级 AI 平台	T8 / T8A 或多节点 T4 / T4A	用户隔离、资源调度、机房条件、运维服务和验收资料
高互联大模型训练、多机扩展、整柜算力	T8X / T-RackScale 项目制方案	NVLink / NVSwitch / IB、供电制冷、整柜交付和项目验收
数据集、素材库、结果归档、快照备份	S / SF 存储产品线	容量、增长率、权限、备份窗口和恢复目标

02. 为什么不能只按 GPU 数量选服务器

常见误区	可能风险
只看 4 卡、8 卡、16 卡	不知道显存是否够，数据是否喂得上，训练效率是否稳定。
只看单卡算力	忽略 GPU 通信、CPU 喂数、存储吞吐和长期满载散热。
只按最低价格配	可能在质保、机房部署、稳定性和交付验证上留下隐患。
把推理服务器当训练服务器	后续做微调、多机训练或多课题组共享时容易遇到瓶颈。
不规划数据路径	GPU 买得很强，但数据集、checkpoint 和共享存储拖慢任务。

03. G 系列和 T 系列的边界

产品线	中文名称	适合场景	关注重点
G 系列	GPU 计算服务器	AI 推理、RAG、本地模型服务、GPU 渲染、视频转码、CAE GPU 加速、多用户 GPU 计算	GPU 吞吐、显存、CPU 喂数、本地 NVMe、网络、稳定性
T 系列	AI 训练服务器	模型训练、微调、多模态、视觉模型、AI 科研、多课题组共享、多机扩展	高显存、GPU 互联、NCCL、数据路径、训练软件栈、满载验证

> **推理 / RAG / 渲染 / 转码 / CAE GPU 加速，优先 G 系列；训练 / 微调 / 多模态 / AI 科研 / 多课题组共享，优先 T 系列。**

04. 2026 年 AI GPU 选型的五个判断顺序

1. 工作负载

先明确推理、微调、训练、视觉、多模态、科学计算和渲染的任务占比。

2. GPU 显存

很多 AI 项目首先遇到的瓶颈不是理论算力，而是单卡显存和总显存。

3. GPU 互联

多卡训练效率受 GPU 互联、NCCL、训练框架、数据路径和模型并行策略影响。

4. 数据路径

本地 NVMe、共享存储、checkpoint、归档备份和网络都会影响 GPU 持续利用率。

5. 机房和交付环境

正式报价前应确认机柜深度、单柜功率、PDU、空调制冷、承重、噪声、远程管理、网络和运维要求。

05. 典型采购场景

企业本地模型推理 / RAG / 多模型服务

小规模试点可从 G2 / G2A 开始；部门级推理平台优先 G4 / G4A；高吞吐、多模型、多用户场景可进入 G8 / G8A。

高校 / 科研院所 AI 微调与多课题组共享

单课题组或实验室节点优先 T4 / T4A；学院级共享平台优先 T8 / T8A；明确需要高互联、多机训练或整柜算力时进入 T8X / T-RackScale。

百万级 AI 科研预算

如果客户预算已经明确，并且用途包括 AI 科研、视觉/多模态、科学计算、多课题组共享，建议按“学院级 AI 科研共享平台”来规划，而不是按普通单台工作站报价。

06. GPU 类型怎么理解

GPU 方向	适合场景
48GB 专业 GPU	推理、视觉模型、小规模微调、渲染、专业图形。
96GB 级专业服务器 GPU	多用户 AI 科研、LLM 微调、多模态、高显存推理、学院级共享平台。
H100 / H200 数据中心 GPU	高端训练、科学计算、大模型微调和训练平台。
B200 / GB200 / 后续 Blackwell 平台	大模型训练、推理工厂、整柜级平台。

> **具体 GPU 型号、供货周期、价格、兼容清单和保修政策，以项目报价时的正式资料为准。**

07. 配置规划必须同时看这些因素

维度	为什么重要	建议关注
CPU	负责数据加载、预处理、请求调度、检索和 PCIe / 网络 / 存储扩展。	不要让 CPU 成为 GPU 喂数瓶颈。
内存	影响数据预处理、容器、缓存和多用户任务。	4 GPU 训练通常从 512GB - 1TB ECC 评估，8 GPU 多用户按项目确认。
存储	影响数据读取、checkpoint、模型文件和结果归档。	本地 NVMe + 共享存储 + 归档备份分层规划。
网络	影响多节点训练、共享存储和推理服务访问。	单机推理 10/25GbE 可起步，训练平台建议评估 100/200/400GbE 或 IB。

08. 交付不只是硬件装好

层级	验证内容
硬件层	GPU 识别、温度、风扇、电源、BMC、NVMe、网卡。
系统层	OS、Driver、CUDA、cuDNN、NCCL、Docker/容器。
训练层	GPU 满载、显存压力、NCCL 通信、训练样例、本地 NVMe 吞吐。
运维层	交付清单、配置档、版本记录、质保说明、故障响应路径。

英睿特的目标不是简单交付设备，而是帮助客户把服务器落到可使用、可验证、可维护的 AI 研发环境。

09. 采购前建议准备的信息

1. 主要任务：推理、训练、微调、多模态、视觉、科学计算、渲染或混合负载；
2. 模型规模和数据规模；
3. 使用人数或课题组数量；
4. 预算范围，是否含税、含运维、需要盖章报价；
5. 机房条件：机柜、供电、制冷、网络、存储；
6. 预计采购时间和交付要求；
7. 是否需要推荐方案和备选方案。

10. 获取配置建议

如果您正在规划 AI GPU 服务器、AI 科研平台或本地推理/训练环境，可以提交项目需求。英睿特会根据模型类型、数据规模、GPU 显存需求、预算、机房条件和交付要求，给出更贴合实际使用的配置建议。

> **建议提交：单位/部门、主要用途、模型或软件、使用人数/课题组、预算范围、是否含税/盖章/运维、机房条件、预计采购时间。**

官网
www.yrtserver.com

方案咨询
sales@yrtserver.com

电话
400-965-8299