模块编号	模块名称
M1	大模型推理服务
M2	Embedding 服务
M3	Rerank 服务
M4	向量数据库（Milvus）
M5	文档解析服务
M6	业务接口服务
M7	负载均衡与网关
M8	日志与监控
M9	对象存储
M10	带宽与安全

用途说明
部署 Deepseek-32B 等大语言模型，用于回答生成
批量生成文本向量，使用 BGE-M3 模型
BGE-Rerank-M3 模型排序
向量存储与相似度检索
OCR / MinerU 文档结构化
API 接入、用户管理
请求调度、安全控制
系统日志、运行指标
文档与向量文件存储
出口带宽、防火墙、WAF

推荐云服务类型
GPU 云主机（高性能）
GPU 云主机（中等）
GPU 云主机（中等）
高性能存储优化型主机
计算型主机（可加速）
计算型云主机
SLB + API 网关
计算型主机
对象存储服务（OBS/OSS）
云基础设施

推荐配置
GPU：8 × A100 80G；CPU：64 vCPU；内存：512 GB；存储：2 TB NVMe SSD
GPU：2 × A100 40G；CPU：32 vCPU；内存：128 GB；存储：1 TB SSD
同上
CPU：32 vCPU；内存：256 GB；存储：4 TB SSD
CPU：32 vCPU；内存：128 GB；可选 GPU：T4/3090
CPU：8 vCPU；内存：32 GB
云服务资源
CPU：8 vCPU；内存：32 GB；存储：2 TB HDD
≥4 TB 存储；≥50MB/s 吞吐
≥200 Mbps 出口

数量	说明备注
1	满足 >50 并发问答，支持多线程推理
1	独立部署优化吞吐
1	可与 M2 合并
1	支持高并发 ANN 查询
1	按需运行，不常驻
2	主备部署
1	支持 HTTPS、安全防护
1	搭配 Prometheus / ELK 使用

- 支持冷热分层
- 满足 50+ 并发需求

弹性扩容建议
可按需增加实例，或使用 vLLM 动态分配线程
可弹性增加至 4 卡或多实例并发处理
如响应压力大可拆分独立部署
可配置分片与副本，或横向扩容节点集群
可与 API 服务共享节点或使用弹性容器运行
使用负载均衡自动分发请求，支持横向扩容
根据 QPS 动态扩容，具备自动伸缩能力
可对接云监控服务，如阿里云 ARMS
访问量大时升级为高频热存储
可动态调整带宽，WAF 可热扩容