# Kboss 2.0

开元云（北京）科技有限公司

2024年7月

# Kboss 1.0 回顾

- 开元云业务云上
  - 供应商产品上云
  - 分销商业务过程上云
  - 客户购买体验上云
  - 供应商分销商结算
  - 业务考核上云

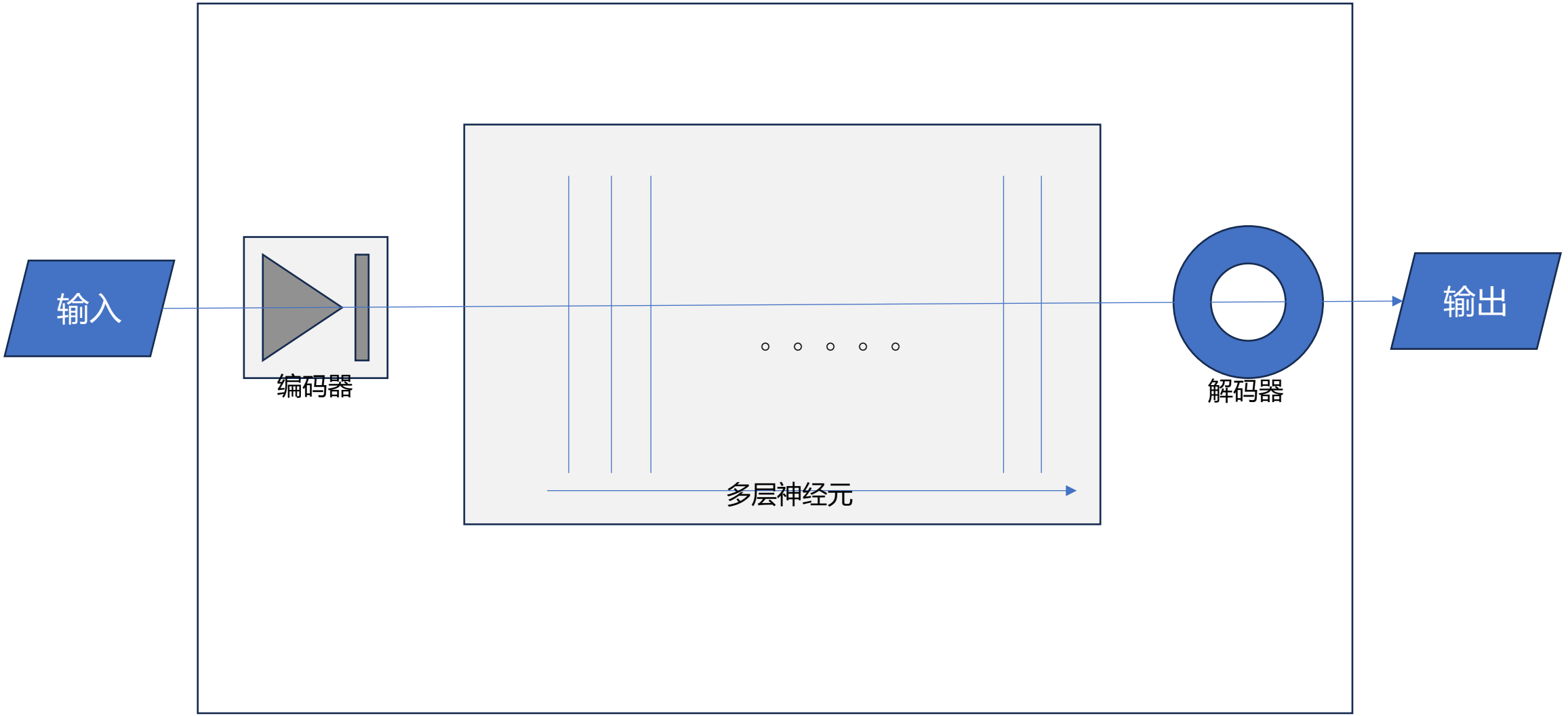
# Kboss 2.0概要

- Kboss平台持续改进
  - 更多资源方接入
  - 服务平台接入
  - MaaS平台接入
- 服务平台
  - 监控与运维
  - 持续集成与部署
  - 安全合规
  - 公共服务部署
- MaaS平台
  - API连接主要LLM，提供客户大模型体验
  - RAG，微调，客制化模型服务
  - 离线或微调模型客制化部署
  - 在线智能体开发，调试，发布
  - 模型Pipeline定义，混合模型应用开发和发布

# 关于大模型和大模型应用

- 什么是大模型
- 从业公司如何分类，都在做什么
- 我们做什么

# 大语言模型（Large Language Model）



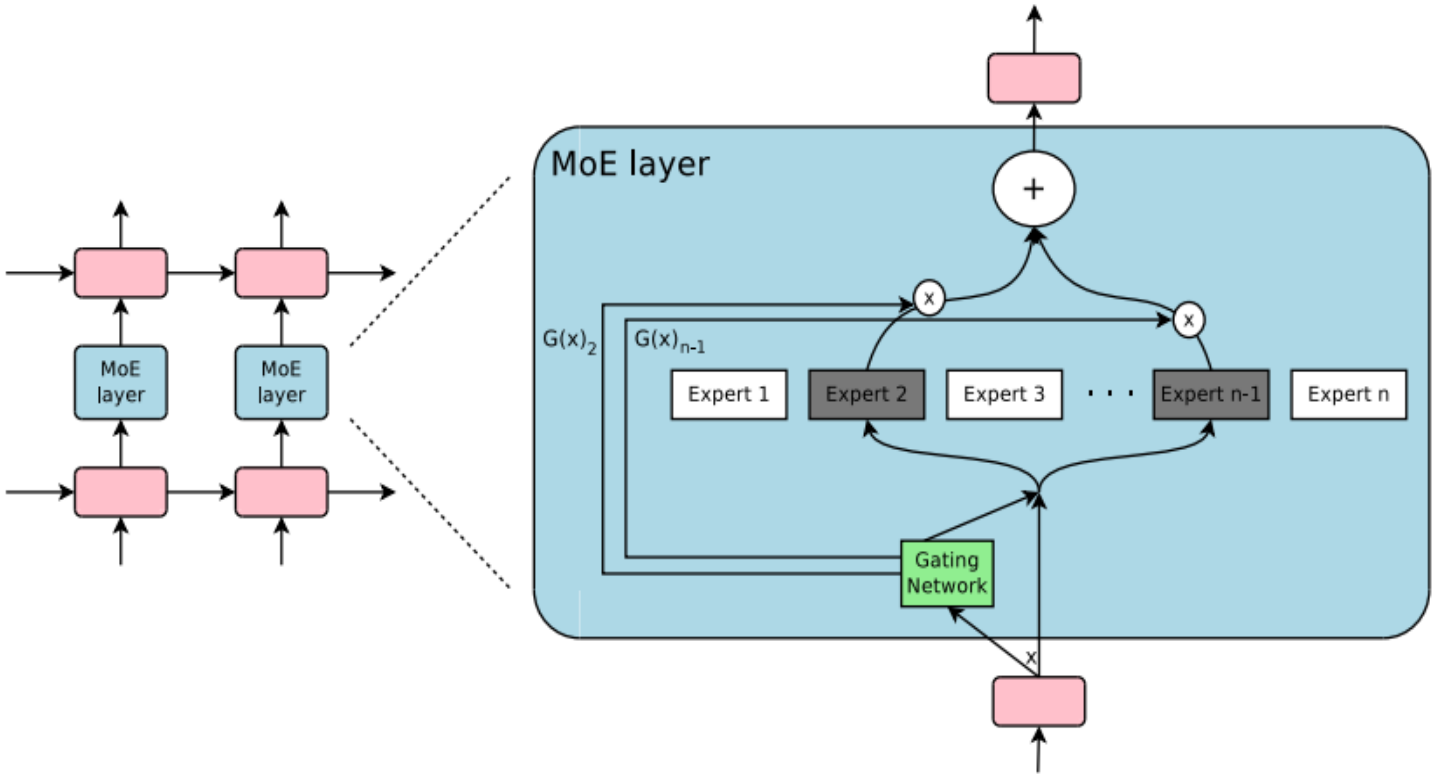
# 模型开发

- 巨量数据准备（Q/A数据）
- 训练环境（万卡规模）
- 长时间训练
- 验证与回归测试
- 定版与发布

- Transformers框架
- 模型规模
- 一个模型可以配置多个编码器和解码器
- 不同的编码器适配不同的输入
- 不同的解码器适配不同的输出

# MoE模型（Mixture of Experts）

- 更快的训练速度
- 更少的GPU需求
- 同等参数量推理速度更快
- 需要更大的显存



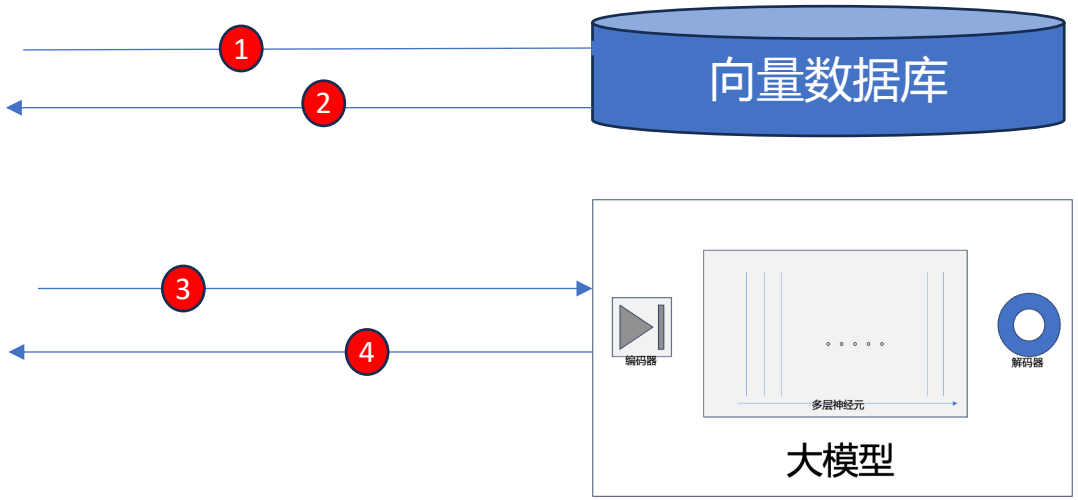
# 模型量化

- **模型量化（Model Quantization）就是通过某种方法将浮点模型转为定点模型**。比如说原来的模型里面的权重（weight）都是float32，通过模型量化，将模型变成权重（weight）都是8bits, 4bits, 2bits或1bits的定点模型
- 模型量化就是建立一种浮点数据和定点数据间的映射关系，使得以较小的精度损失代价获得了较大的收益

| 量化前的浮点模型 | 量化后的定点模型 |
| --- | --- |
| 参数量大（float32） | 压缩参数（int8） |
| 计算量大 | 提升速度 |
| 内存占用多 | 内存占用少 |
| 精度高 | 精度损失 |

# RAG(Retrieval-Augmented Generation)

- **搜索增强生成**（大语言模型体外知识库）
- 向量数据库
- 生成知识库：领域知识的正文文本一次性转化为高维向量数据，并保存在向量数据库中
- 知识检索：问题文本转化为向量，并在知识库中检索空间距离最短的向量，逆向转化为文本返回，通常按具体从小大大排序返回几个结果



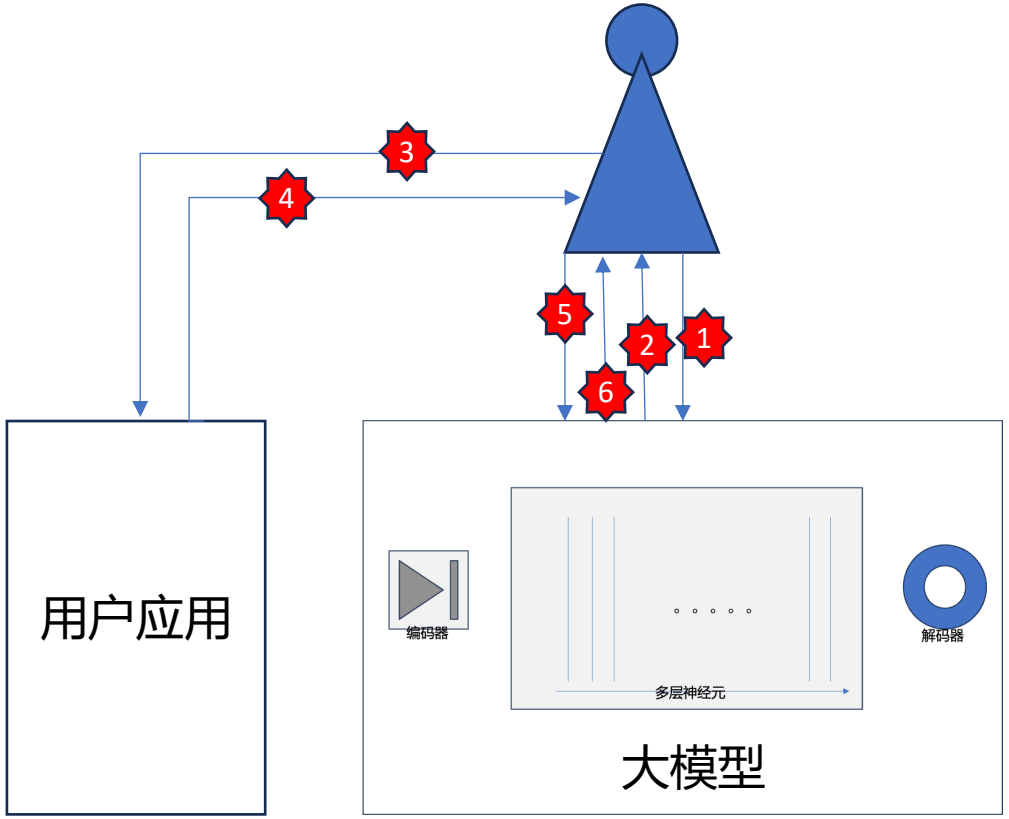

1. 用户输入问题
2. RAG返回文本
3. 用户输入问题 + RAG返回文本
4. 最终应答

# 模型微调（fine tuning）

- 专业知识文本准备（Q/A格式文本）
- 选择模型微调工具
- 选择基座模型（参数越多需要的算力越大，微调时间越长）
- 执行微调
- 验证与回归测试
- 定版与发布

# 函数调用

- 作用：将用户的一句话翻译成一个预先定义的函数集中的一个函数，json格式以函数名，所需参数名称和值的形式返回。
- 大模型不执行 所定义的函数
- 用户需实现将函数按照大模型约定格式要求预先定义函数，可以定义多个函数
- 需要体外设置函数识别规则，并有能力执行函数




1. 函数定义 + 用户输入
2. 函数名，函数参数
3. API调用
4. 返回调用结果
5. 函数定义+调用结果
6. 返回应答

# LLM总结

- 大语言模型（稠密模型，主要性能指标：参数规模）
- MOE模型（稀疏模型，主要性能指标：参数规模）
- 量化模型（小体量模型，主要性能指标：参数规模，位数）
- 微调模型（在基座模型上用特定知识文本训练后的模型，主要性能指标与基座模型相同）
- RAG（模型体外知识库）
- 函数调用（大模型需理解函数，需外部系统执行函数）

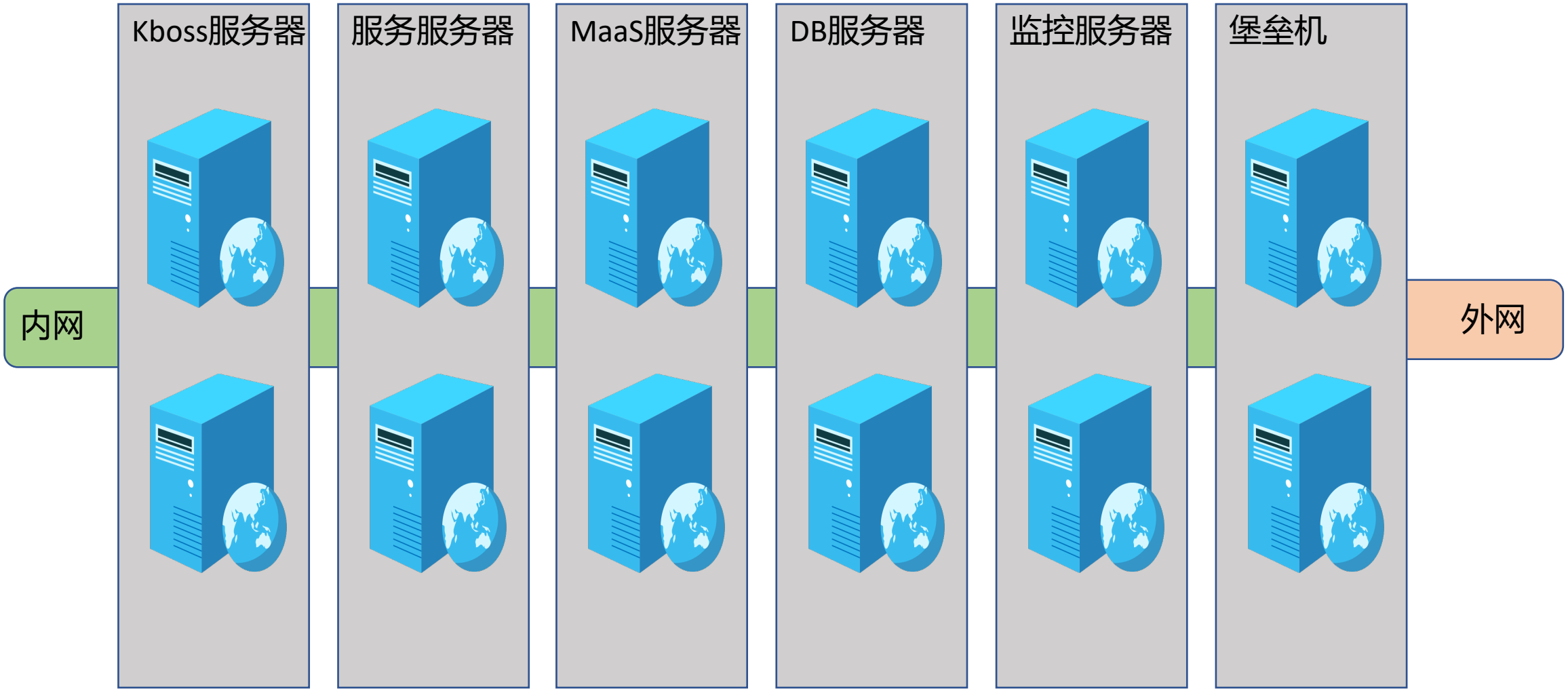
# 从业分类与专注

- 大模型公司
  - 生产大模型
- 平台公司
  - 融合大模型平台，
  - 提供agent，角色扮演
- 应用商店
  - 应用上架展示
  - 在线购买
  - 下载
- 应用开发
  - 行业大模型微调，知识库以及应用集成
  - 功能性应用开发

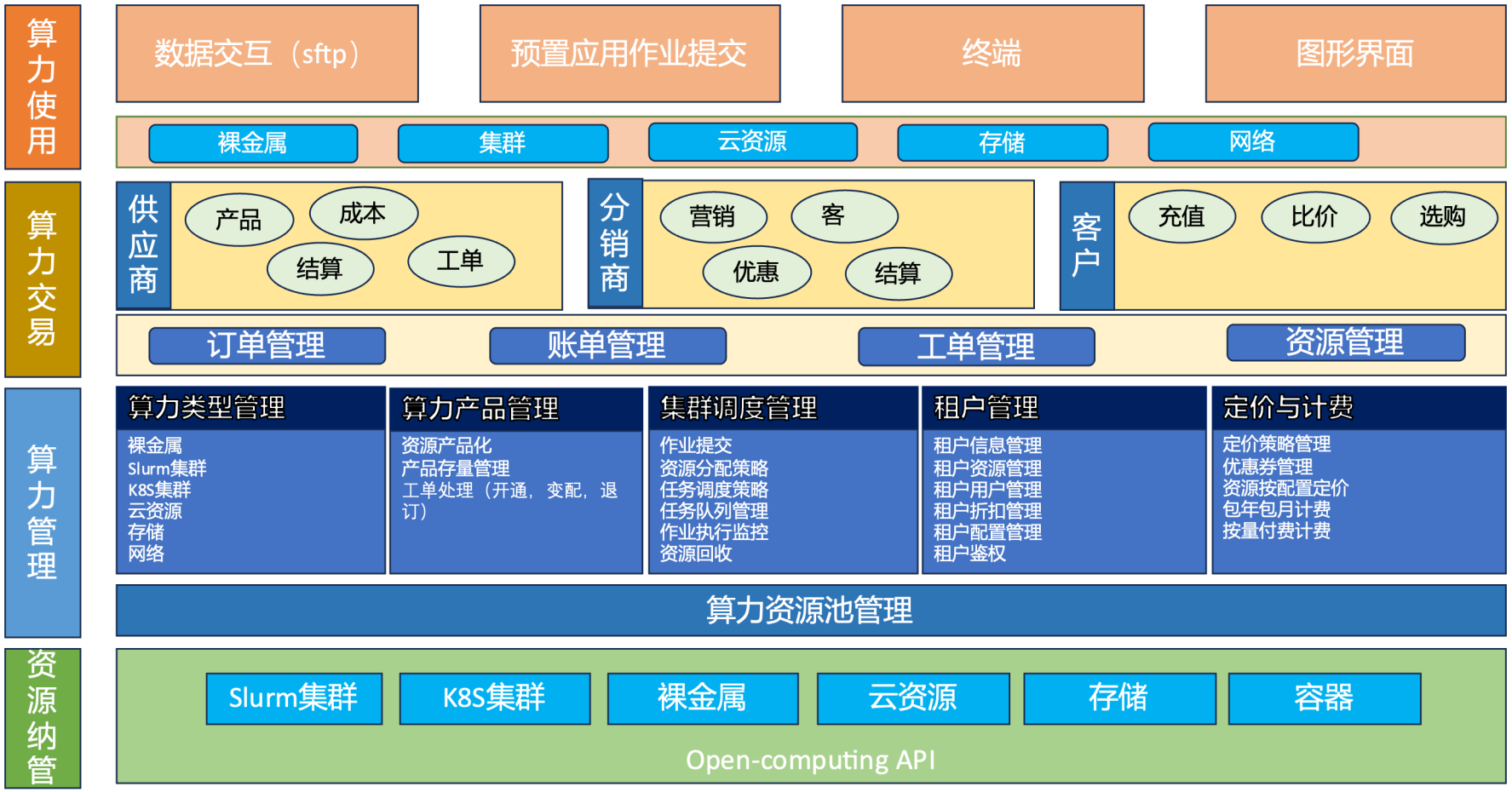
# 我们做什么

- 大模型平台（MaaS）
- 应用商店（为Kboss2.0提供大模型应用产品）
- 大模型行业应用
- 大模型功能性应用开发

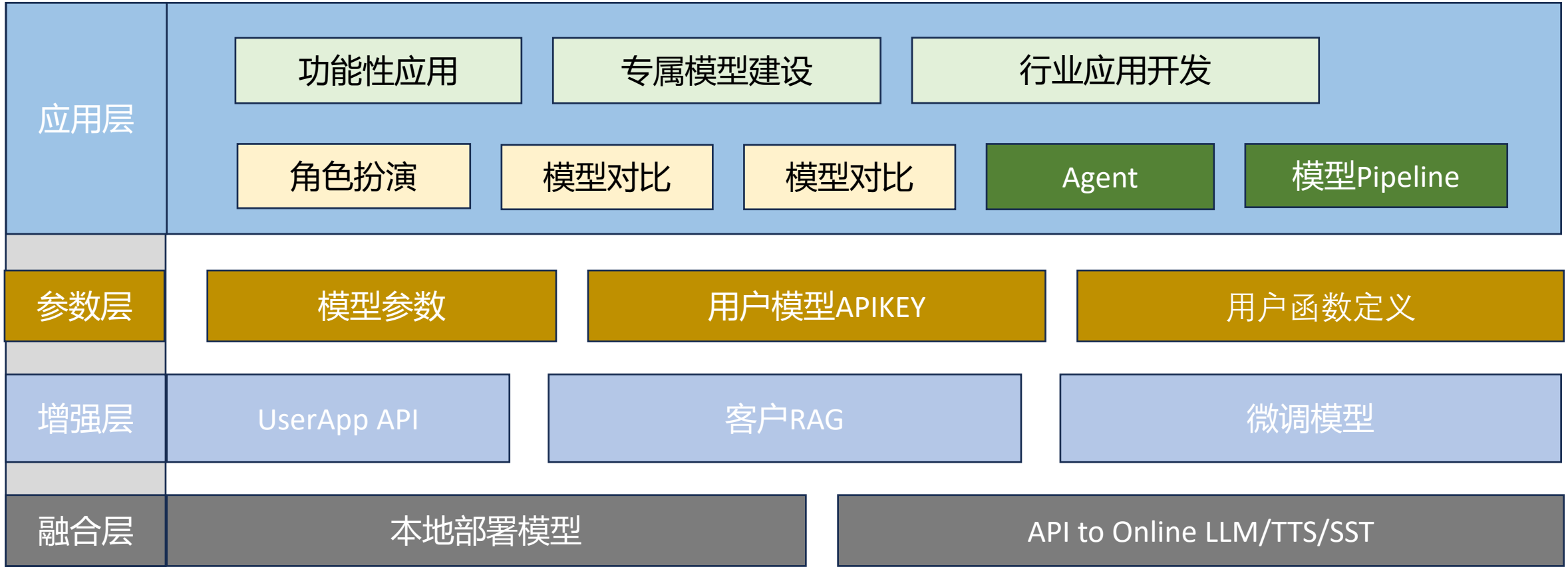
# Kboss 2.0部署架构



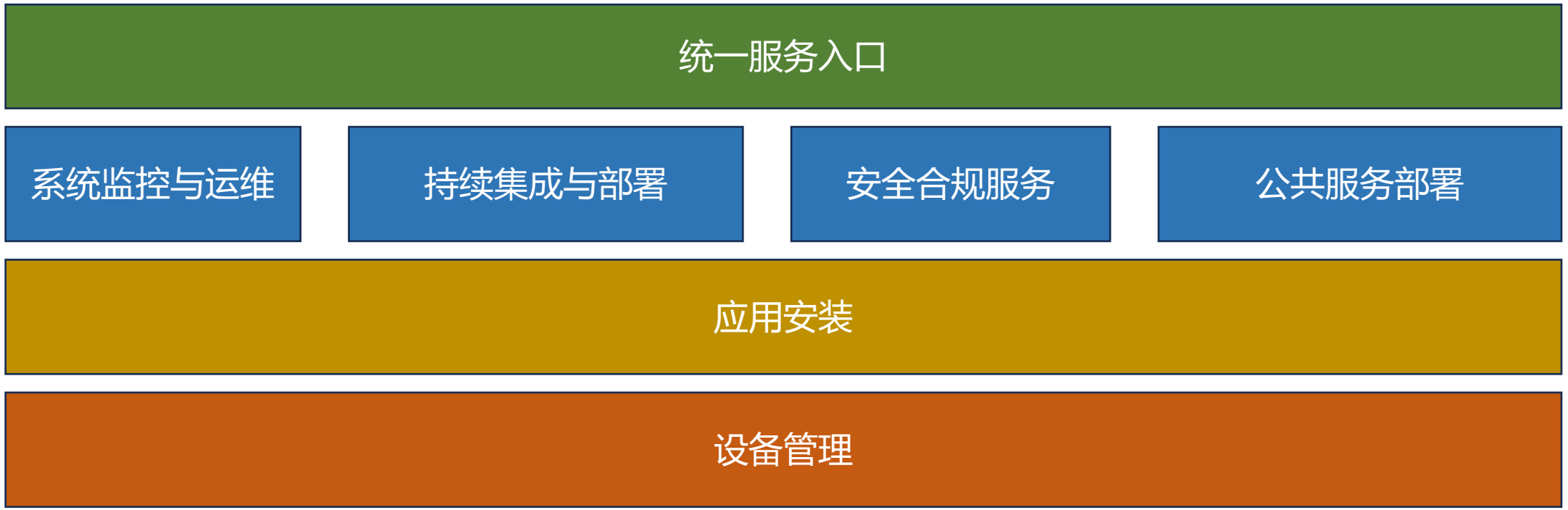
# Kboss平台架构


算力使用
数据交互（sftp）
预置应用作业提交
终端
图形界面
裸金属
集群
云资源
存储
网络
算力交易
供应商
产品
成本
结算
工单
分销商
营销
客
优惠
结算
客户
充值
比价
选购
订单管理
账单管理
工单管理
资源管理
算力管理
算力类型管理
裸金属
Slurm集群
K8S集群
云资源
存储
网络
算力产品管理
资源产品化
产品存量管理
工单处理（开通，变配，退订）
集群调度管理
作业提交
资源分配策略
任务调度策略
任务队列管理
作业执行监控
资源回收
租户管理
租户信息管理
租户资源管理
租户用户管理
租户折扣管理
租户配置管理
租户鉴权
定价与计费
定价策略管理
优惠券管理
资源按配置定价
包年包月计费
按量付费计费
算力资源池管理
资源纳管
Slurm集群
K8S集群
裸金属
云资源
存储
容器
Open-computing API

# MaaS平台架构


应用层
功能性应用
专属模型建设
行业应用开发
角色扮演
模型对比
模型对比
Agent
模型Pipeline
参数层
模型参数
用户模型APIKEY
用户函数定义
增强层
UserApp API
客户RAG
微调模型
融合层
本地部署模型
API to Online LLM/TTS/SST

# 服务平台架构



# Kboss平台持续改进

- 更多的资源方接入
  - 阿里云， AWS， 首都云， 火山云， ucloud， 算法互联，。。。
- 独立分销部署
  - 独立的软件部署
  - 与Kboss之间API链接
  - 客户支付给分销商
  - 与开元云定期结算
- 服务平台作为资源方接入
- MaaS平台作为资源方接入
- 基于大模型的比价与推荐

# MaaS平台

- 与Kboss平台对接，作为资源方提供下列产品
  - 大模型对话式体验
  - 客制化模型独立部署（含硬件客制化，模型客制化）
  - 智能体
  - 混合模型应用
- 定位：模型应用开发平台
  - 联通主流在线模型
  - 部署主流离线模型
  - 在线开发调试部署外部应用连接
  - 提供知识库，模型微调开发，测试，部署
  - 智能体开发，测试和部署
  - 混合模型应用开发，测试和部署

# MaaS平台开发里程碑

- 国内主要文生文在线模型联通 - 已完成
- 在线模型用户体验上线 - 8月10日
- Kboss大模型体验产品接口开通 - 8月10日
- 国内主要文生图在线模型联通 - 9月10日
- 国内主要文生视频在线模型 - 9月20日
- 国内主流短视频厂商上传视频接口开通 - 10月20日
- 短视频生成发布平台上线 - 11月30日
- 企业私有化知识库服务搭建 - 12月30日

# 需销售市场部门协调办理

- 代理协议
  - 阿里千问
  - 百度千帆
  - 百川
  - 智普
  - Minimax（已确定）
  - 月之暗面
  - 火山豆包
  - deepseek

- 行业应用需求
  - 高校

# 服务平台

- 与Kboss集成，作为资源方提供
  - 监控和运维服务
  - 持续集成与部署
  - 安全合规
  - 公共服务部署
- 定位：在线应用服务
  - 定制化应用监控与运维
  - 支持git的CI/CD
  - 常用服务的一键式部署
  - 安全与合规服务