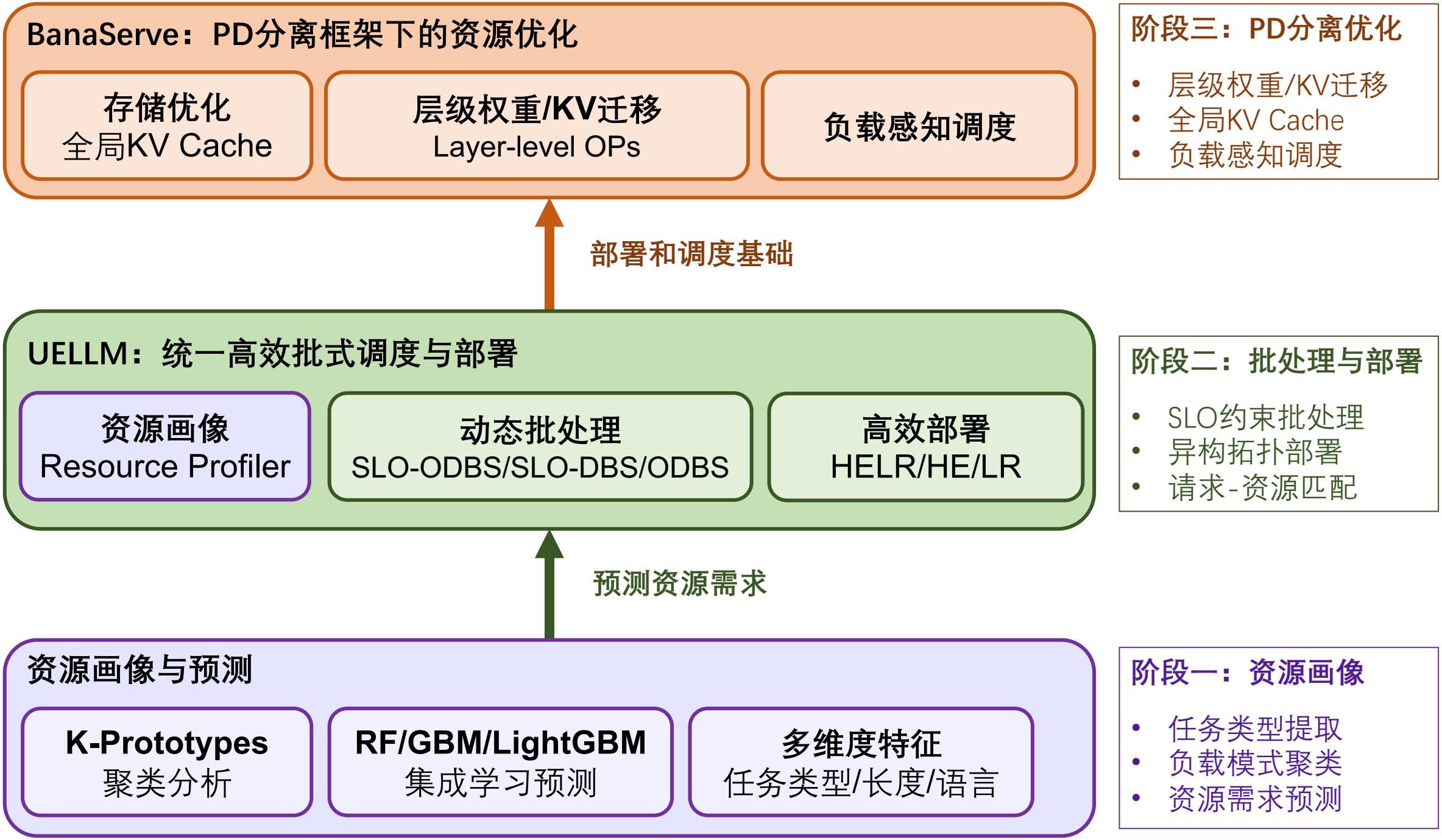
BanaServe：PD分离框架下的资源优化
存储优化
全局KV Cache
层级权重/KV迁移
Layer-level OPs
负载感知调度
部署和调度基础
UELLM：统一高效批式调度与部署
资源画像
Resource Profiler
动态批处理
SLO-ODBS/SLO-DBS/ODBS
高效部署
HELR/HE/LR
预测资源需求
资源画像与预测
K-Prototypes
聚类分析
RF/GBM/LightGBM
集成学习预测
多维度特征
任务类型/长度/语言
阶段三：PD分离优化
层级权重/KV迁移
全局KV Cache
负载感知调度
阶段二：批处理与部署
SLO约束批处理
异构拓扑部署
请求-资源匹配
阶段一：资源画像
任务类型提取
负载模式聚类
资源需求预测