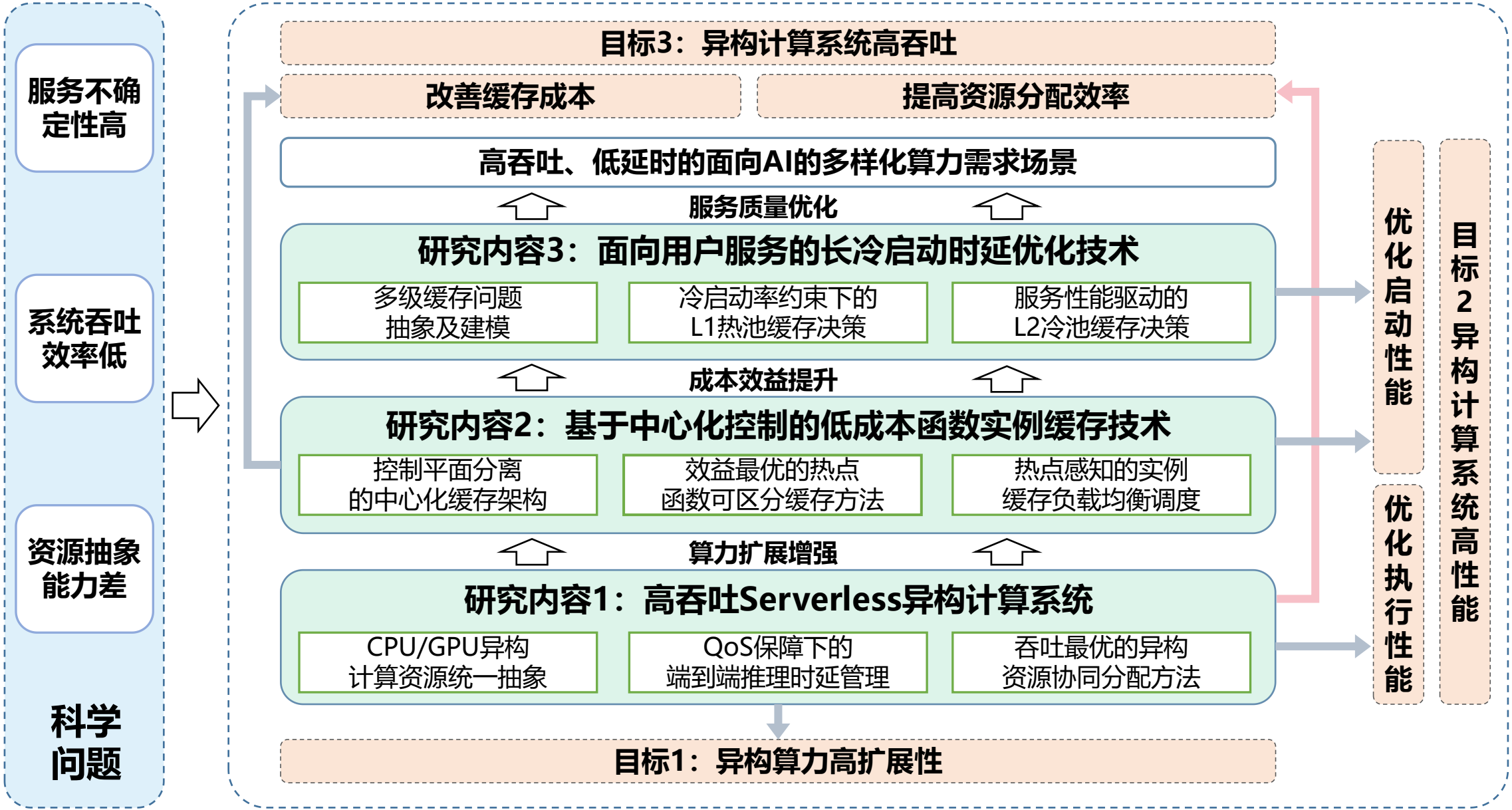
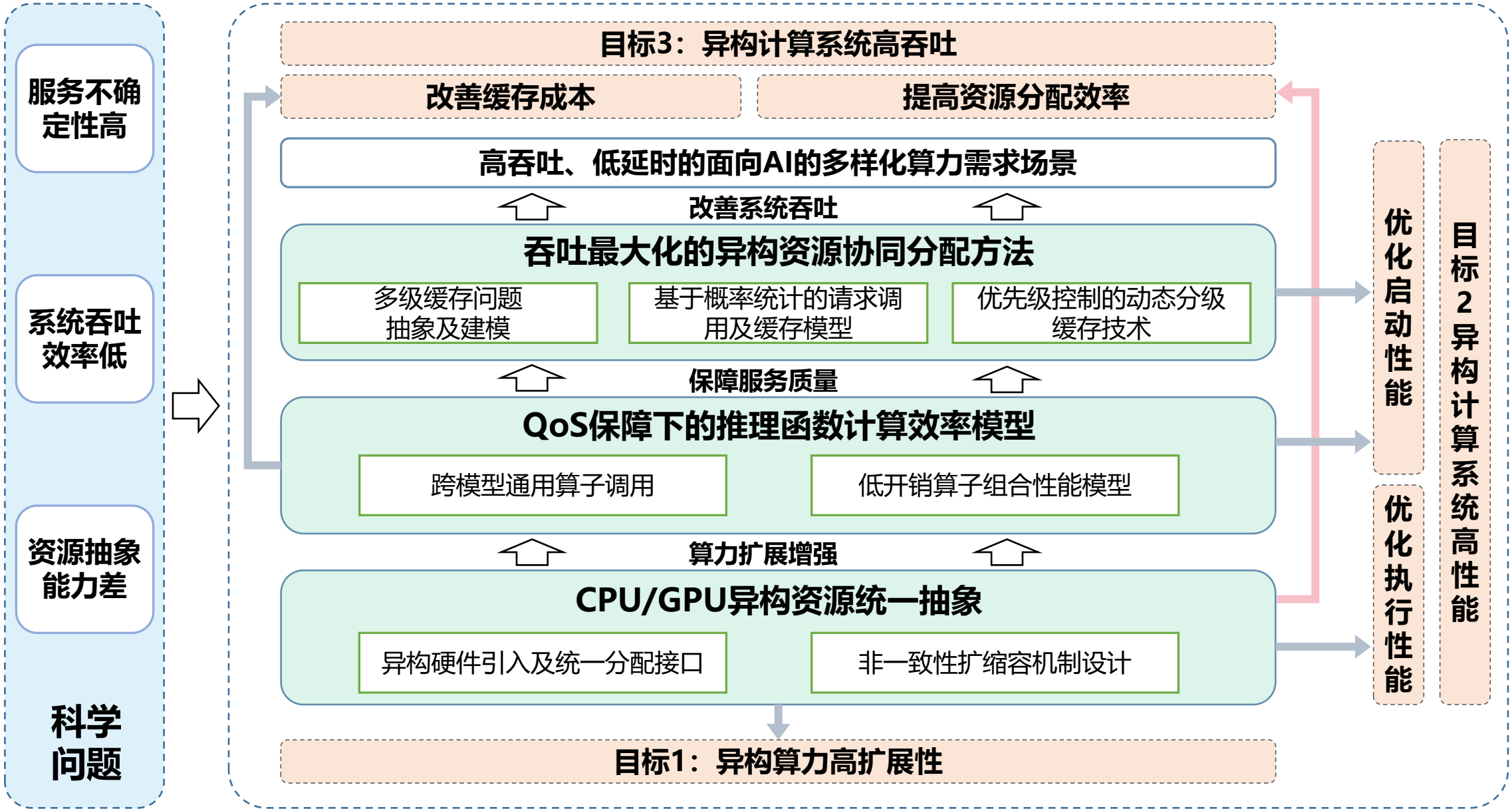

服务不确定性高
系统吞吐效率低
资源抽象能力差
科学问题
目标3：异构计算系统高吞吐
改善缓存成本
提高资源分配效率
高吞吐、低延时的面向AI的多样化算力需求场景
服务质量优化
研究内容3：面向用户服务的长冷启动时延优化技术
多级缓存问题抽象及建模
冷启动率约束下的L1热池缓存决策
服务性能驱动的L2冷池缓存决策
成本效益提升
研究内容2：基于中心化控制的低成本函数实例缓存技术
控制平面分离的中心化缓存架构
效益最优的热点函数可区分缓存方法
热点感知的实例缓存负载均衡调度
算力扩展增强
研究内容1：高吞吐Serverless异构计算系统
CPU/GPU异构计算资源统一抽象
QoS保障下的端到端推理时延管理
吞吐最优的异构资源协同分配方法
目标1：异构算力高扩展性
优化启动性能
优化执行性能
目标2异构计算系统高性能

研究内容2：基于中心化控制的低成本函数缓存技术

研究内容1：高吞吐Serverless异构计算系统


服务不确定性高
系统吞吐效率低
资源抽象能力差
科学问题
目标3：异构计算系统高吞吐
改善缓存成本
提高资源分配效率
高吞吐、低延时的面向AI的多样化算力需求场景
改善系统吞吐
吞吐最大化的异构资源协同分配方法
多级缓存问题抽象及建模
基于概率统计的请求调用及缓存模型
优先级控制的动态分级缓存技术
保障服务质量
QoS保障下的推理函数计算效率模型
跨模型通用算子调用
低开销算子组合性能模型
算力扩展增强
CPU/GPU异构资源统一抽象
异构硬件引入及统一分配接口
非一致性扩缩容机制设计
目标1：异构算力高扩展性
优化启动性能
优化执行性能
目标2异构计算系统高性能