# 京东阿基米德

**对于数据中心资源的单纯静态划分和使用方式的重新考量**

TIG-集群技术部
鲍永成

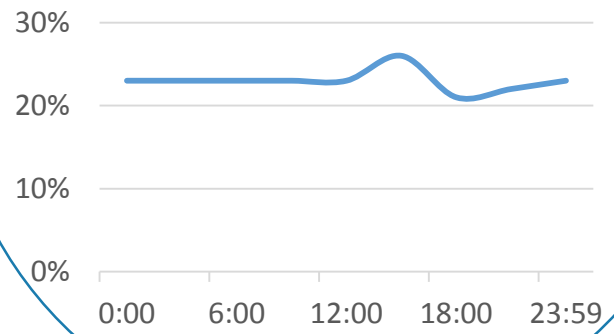
# Agenda

- ✓ 数据中心现状
- ✓ 阿基米德架构
- ✓ 京东容器技术生态
- ✓ 应用画像
- ✓ 抢占式调度
- ✓ 不止于调度

# 资源利用率现状

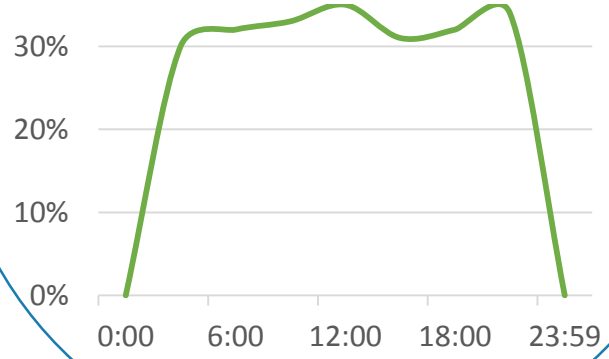
1

## 在线数据资源使用率



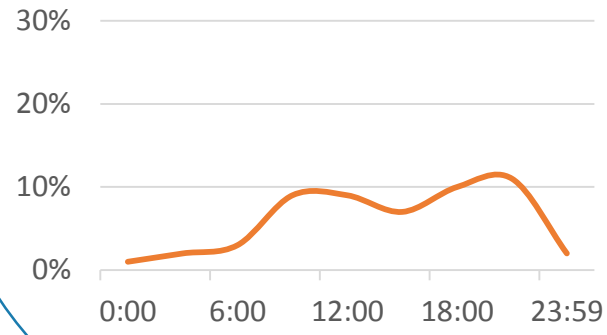

2

## 离线计算资源使用率

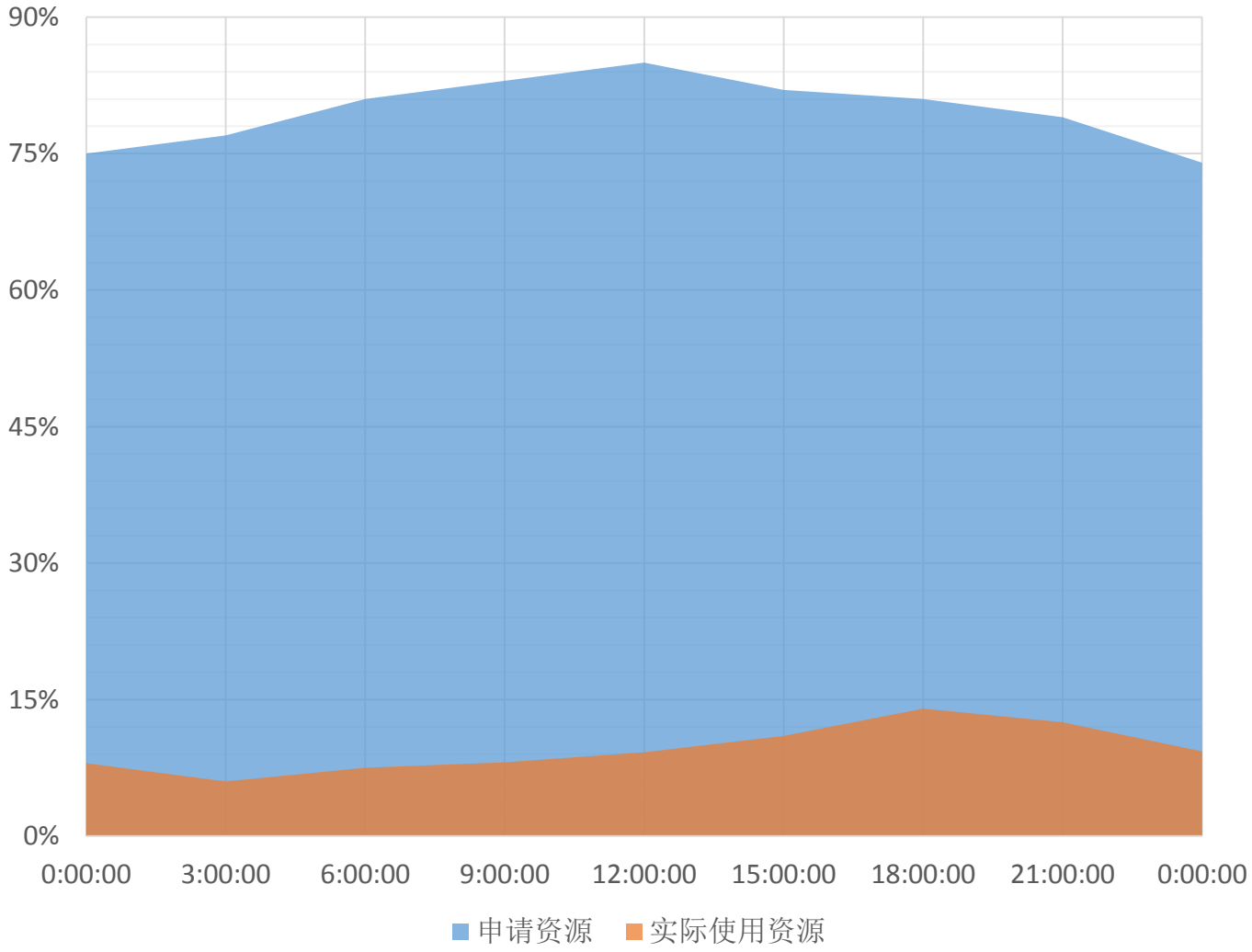



3

## 在线业务系统资源使用率



# 应用资源申请量与实际使用量间差距巨大



# 实现IDC级别调度的难点

### 01.异构问题

服务器的异构

资源的异构(SSD/GPU等特殊资源)

### 02.隔离问题

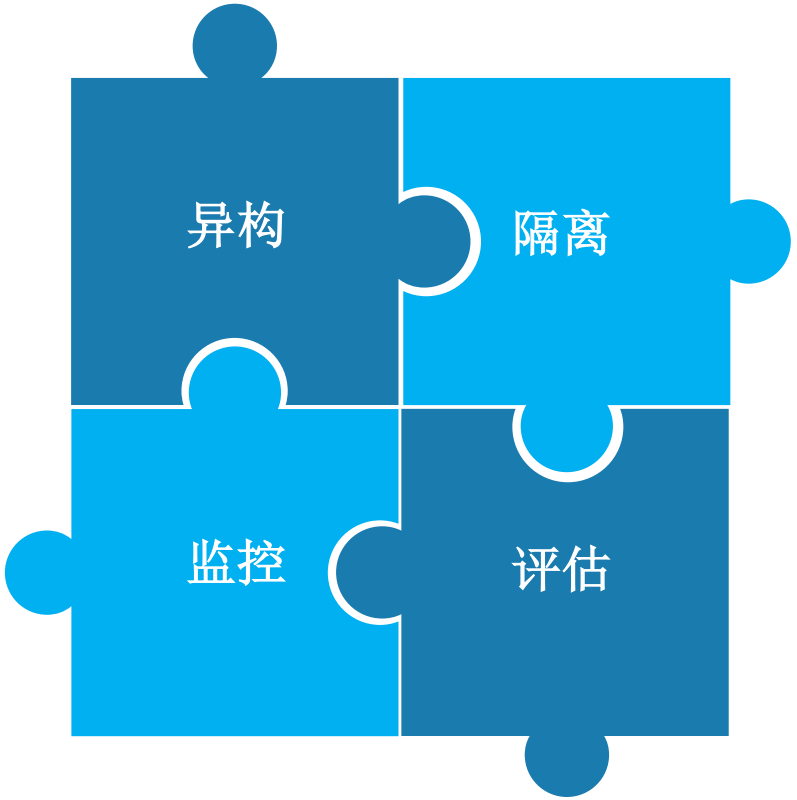
任务之间相互隔离，不互相影响

任务使用的资源能够做有效限制

### 03.监控问题

多方位监控指标提供调度参考

丰富的历史监控数据分析

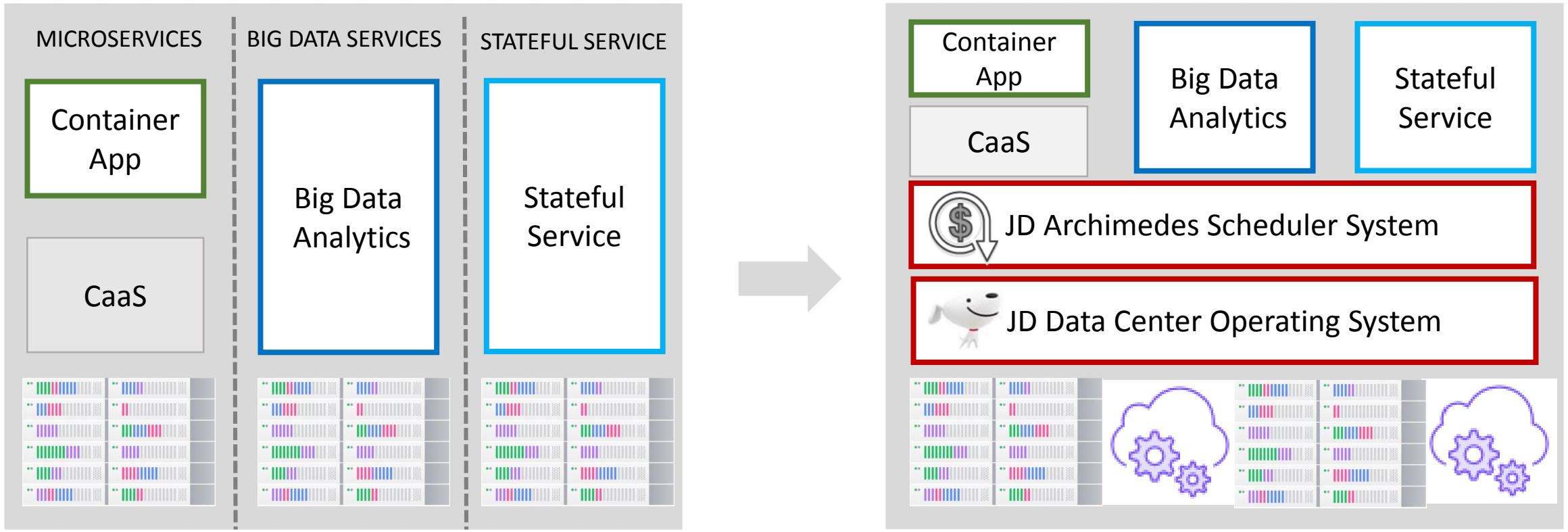



### 04.评估问题

对任务所需资源进行评估

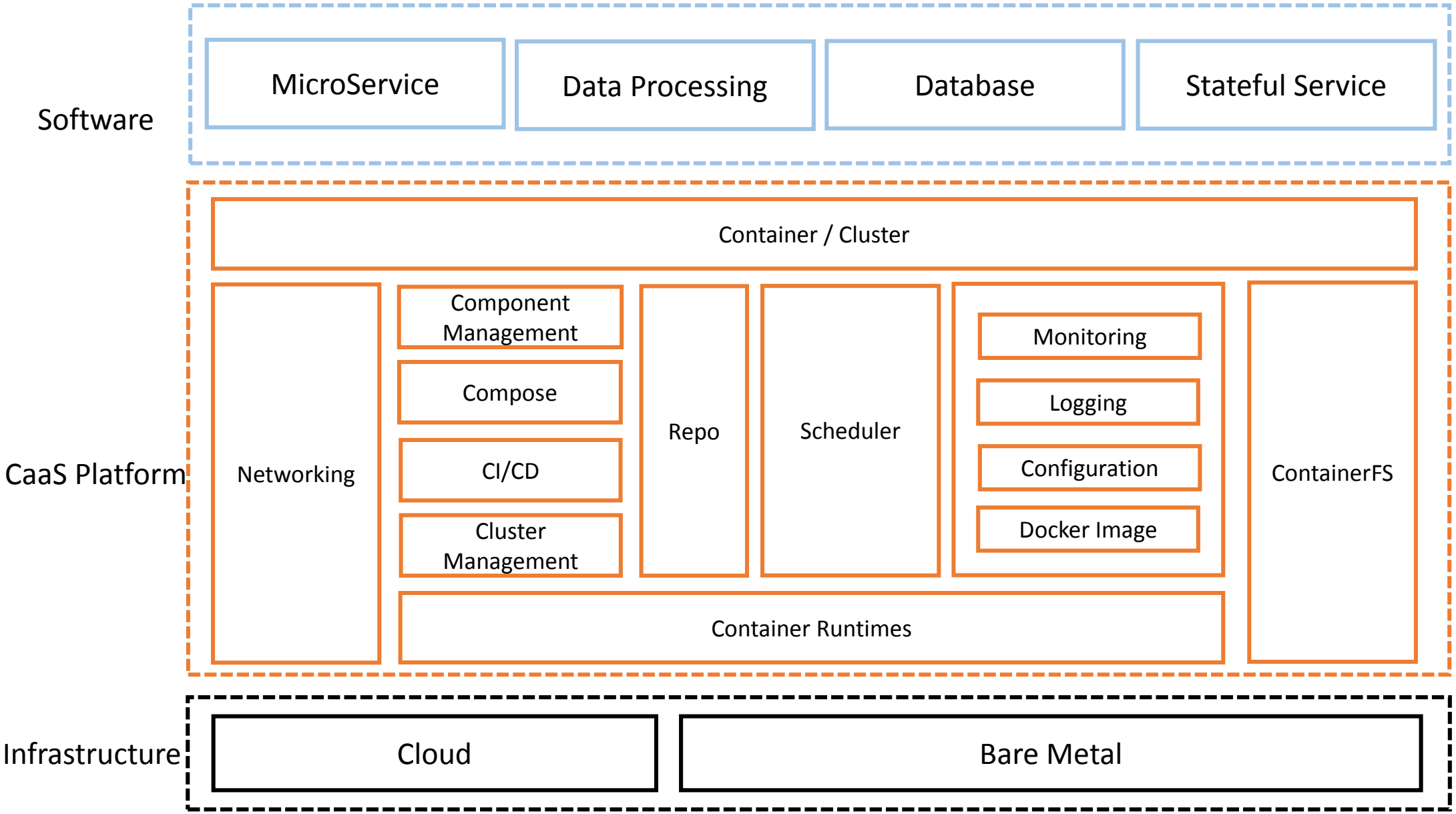
对服务器负载进行评估

# Datacenter Cluster Management & Resource Scheduling




- ✓ **Unified Hybrid Cloud Computing**
- ✓ **Hydrid Deployment and Intelligent Computational**
- ✓ **Lower TCO**

# Container as a Service JDOS

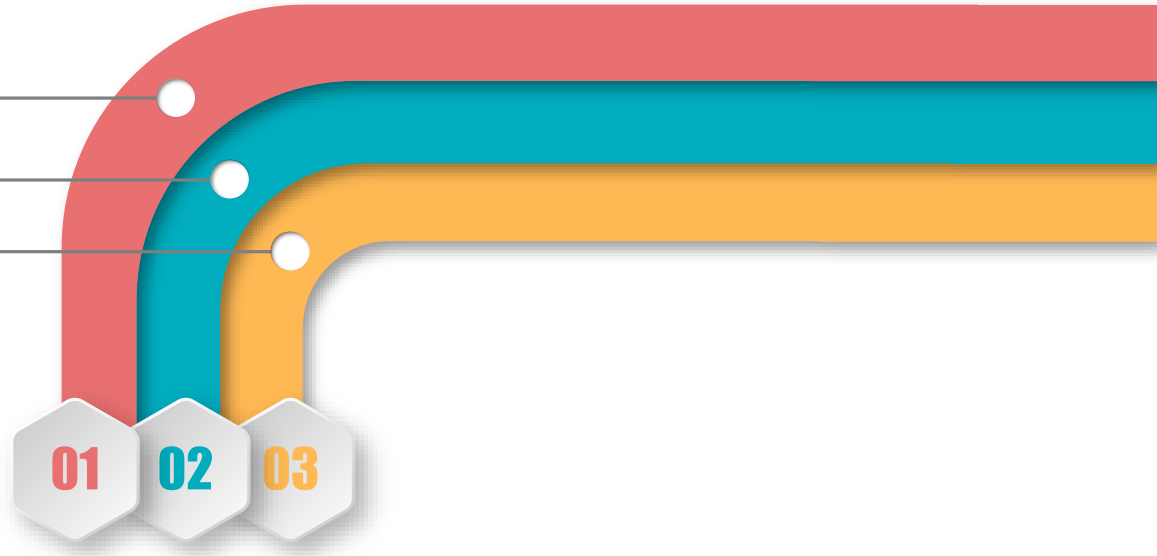


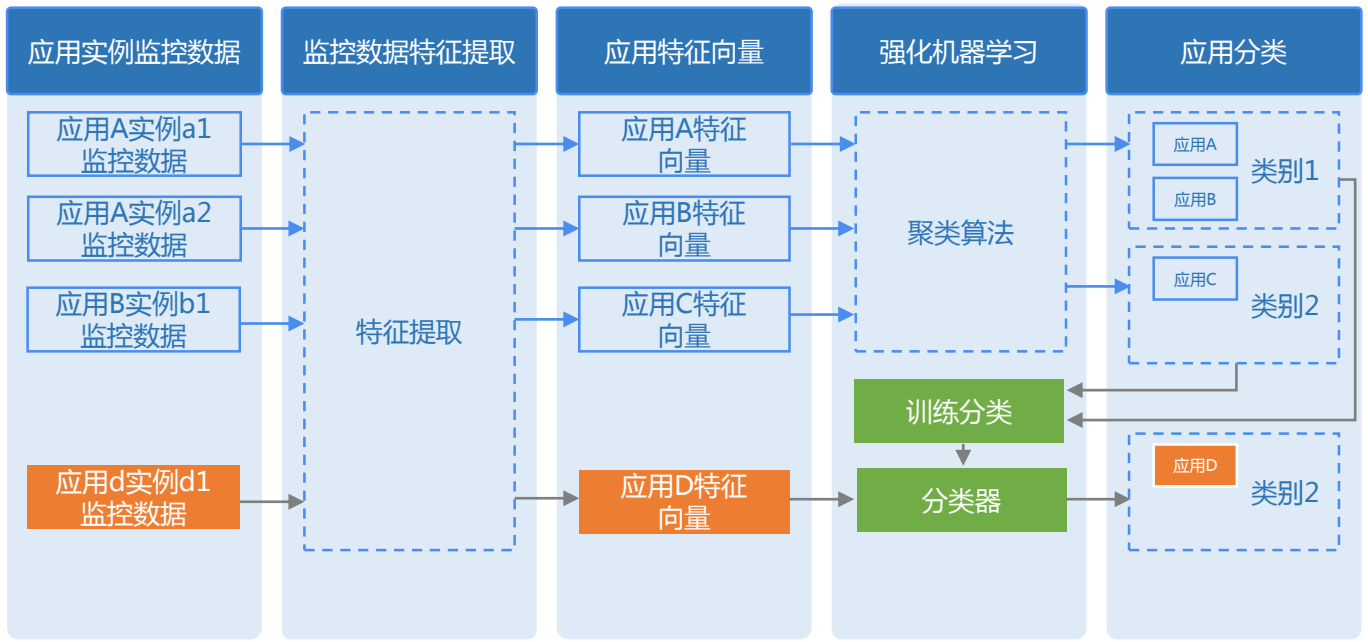
# 应用画像

应用优先级

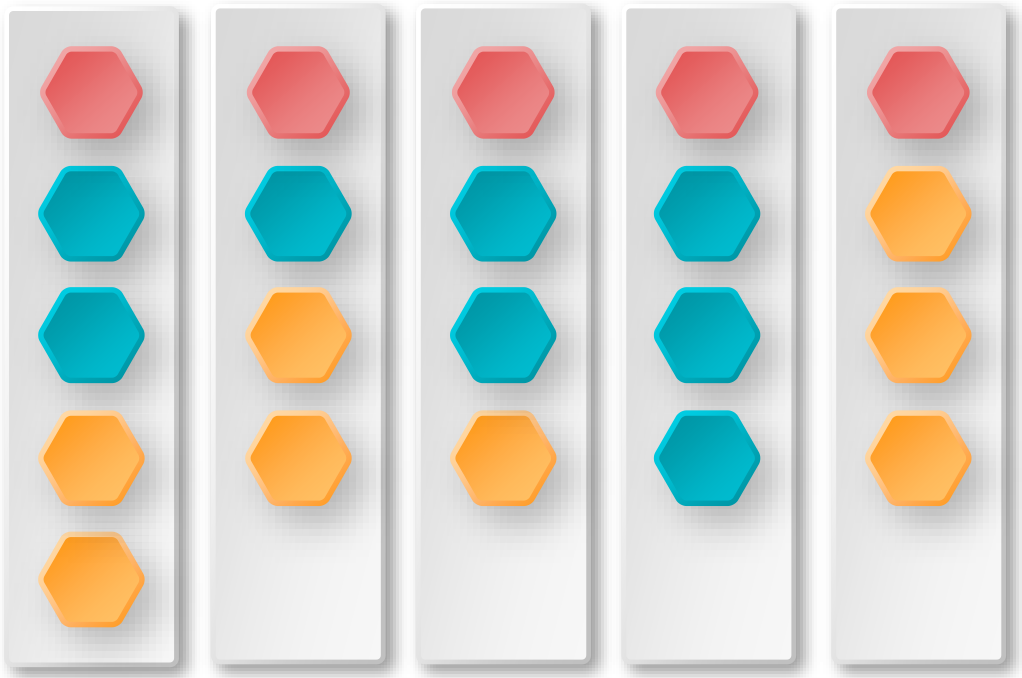
资源使用特征

应用其他描述



# 基于应用画像的调度



高优先级长期服务

低优先级长期服务

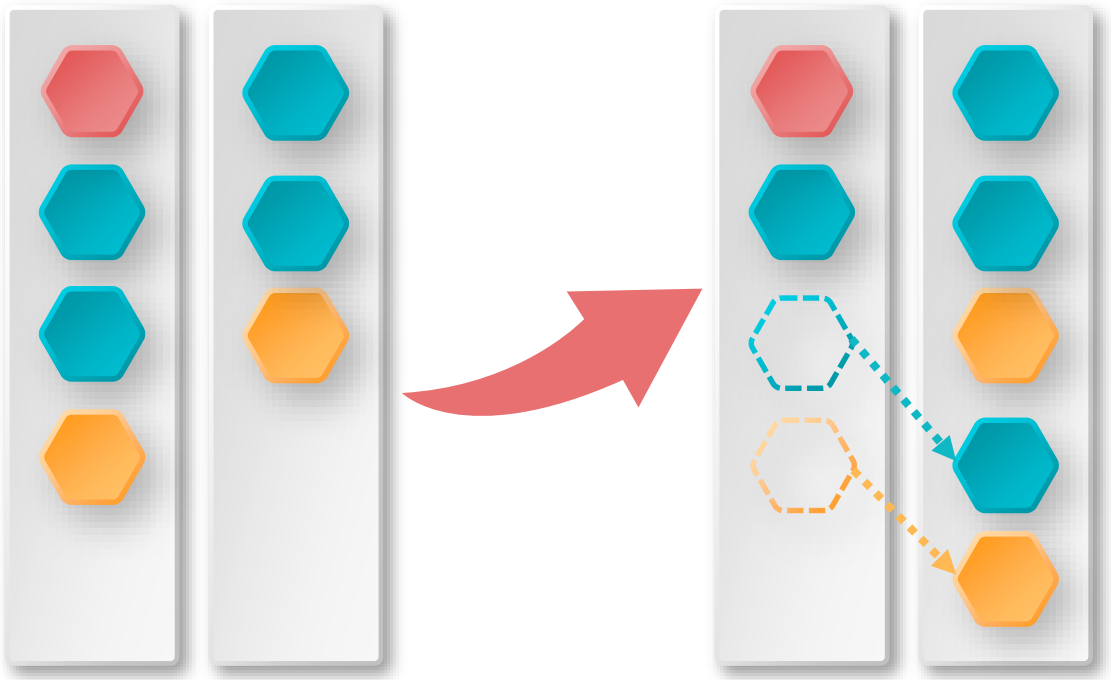
批处理任务

1 高优先级任务优先分散部署

2 批处理任务与长期任务混合部署

3 应用资源使用特征相近的服务进行反亲和性调度

# 抢占式调度



除0级(最高优先级)服务外，其他所有服务/任务均需进入抢占式调度算法的计算范围

优先保证高优先级服务的SLO
优先驱逐批处理任务

抢占式调度采用预防式，在资源进入警戒值之后，在资源枯竭之前即进入抢占式调度流程

资源被抢占的服务/任务将重新进入调度队列，在其他合适节点进行执行

发生抢占式调度的节点将被暂时锁定，以防其他服务/任务调度进入

# 核心系统资源保护

## 优先调度

当有核心系统容器进入调度队列后
将优先为其进行资源调度

## 预留集群资源buffer

为核心系统保留一定的资源buffer
以备其随时进行横向扩展

## 内核级资源保护

特别设置cpushare/oom_adj等值，保证
容器优先获取CPU资源，最迟发生OOM

## 抢占式调度

在发生资源枯竭之前
将其他服务/任务进行驱逐

# 不止于调度

项目管理

统一管理所有资源

显著提高使用密度

项目成本预算/核算

采购部门

科学规划配置升级

降低硬件采购成本

采购计划决策支持

审计部门

研发资产核算明确

研发成本精准透明

资源使用审计

业务部门

重新定义部署方式

融合调度计算资源

应用扩容缩容决策支持