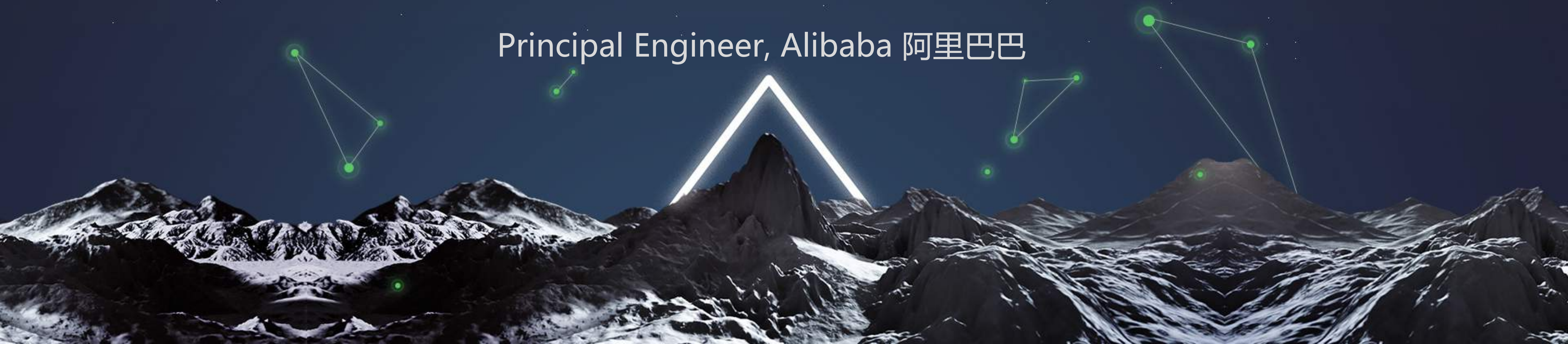


阿里巴巴调度与集群管理系统 Sigma

Liping Zhang
张瓌坪

Principal Engineer, Alibaba 阿里巴巴



QCon

全球软件开发大会

成为软件技术专家的 必经之路

[北京站] 2018

2018年4月20-22日 北京·国际会议中心

7折 购票中, 每张立减2040元
团购享受更多优惠



识别二维码了解更多



极客时间

重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新



扫一扫下载极客时间App

AiCon

全球人工智能与机器学习技术大会

助力人工智能落地

2018.1.13 - 1.14 北京国际会议中心



扫描关注大会官网

SPEAKER INTRODUCE



Liping Zhang 张瓌坪

Alibaba Principal Engineer

阿里花名谷朴。2017年加入阿里巴巴。阿里巴巴系统软件部研究员，调度系统首席架构师。

之前在Google的基础设施事业群的集群管理部门工作了5年多，并领导了资源管理和优化调度团队，负责Borg以及基础存储资源的优化，负责了FlexBorg, Autoscaling等多个产品。加入Google前在加州大学伯克利分校从事智能系统的研究工作。本科和博士毕业于清华大学。

调度系统业界现状

- 开源领域：Kubernetes , Docker Swarm, Mesos, Yarn等
- 闭源，基于开源的改造或者开源兼容系统：Borg , Fuxi , Sigma等
- 在线和离线任务混部：已知主要是Borg
- 资源利用率状况：Borg领先

真正的挑战？在实际工程中落地，实现资源效率和运维效率的提升

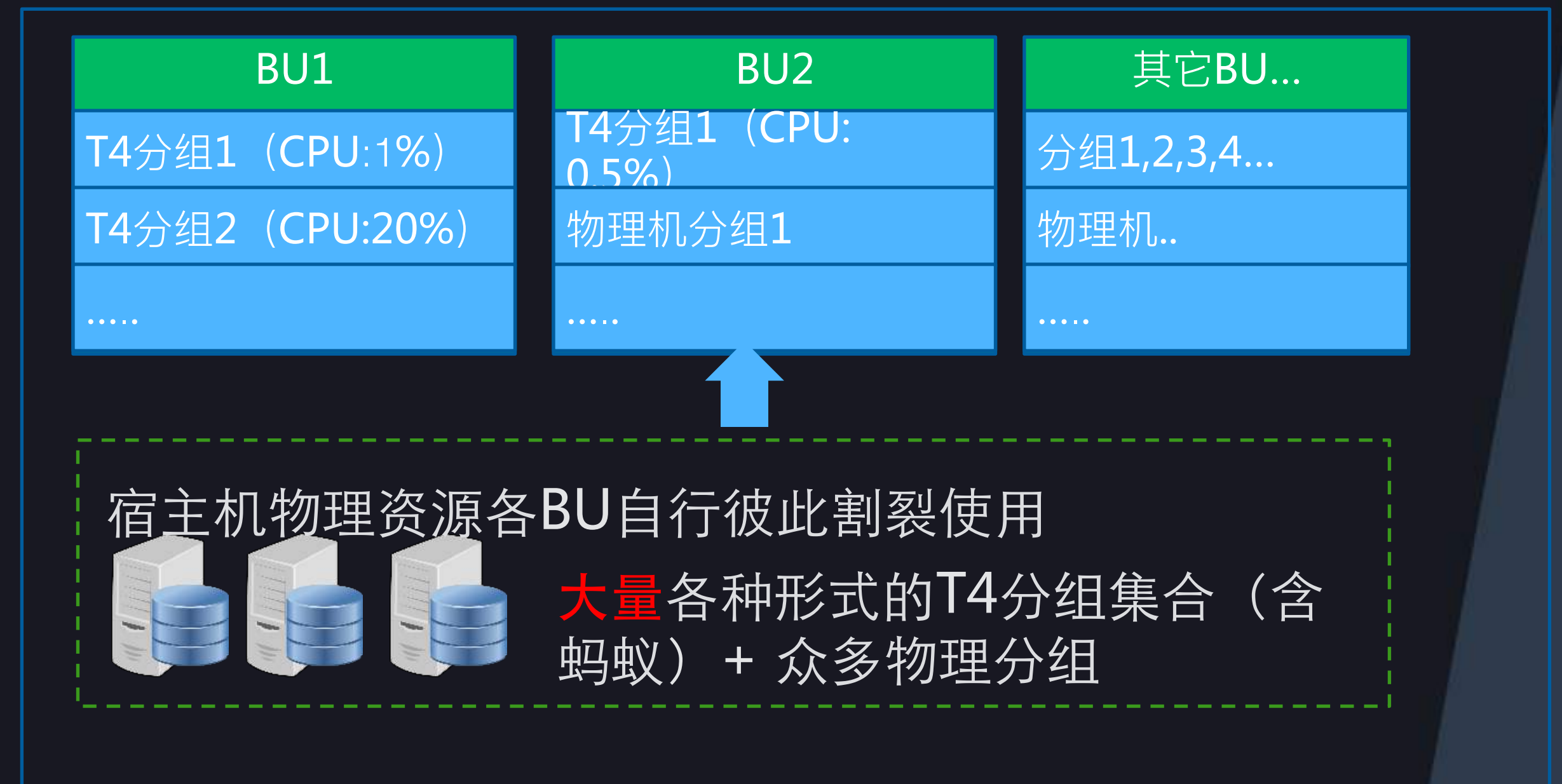
TABLE OF CONTENTS 大纲

- Sigma调度系统整体架构和历史演进
- 架构和实现的特点
- 调度及资源优化
- 展望

阿里调度系统整体架构和历史演进

- 历史状况
 - 各个部门资源池独立
 - 独立研发，多套调度系统
 - 容器演进：T4->Pouch

历史状况



阿里调度系统整体架构和历史演进

●问题：

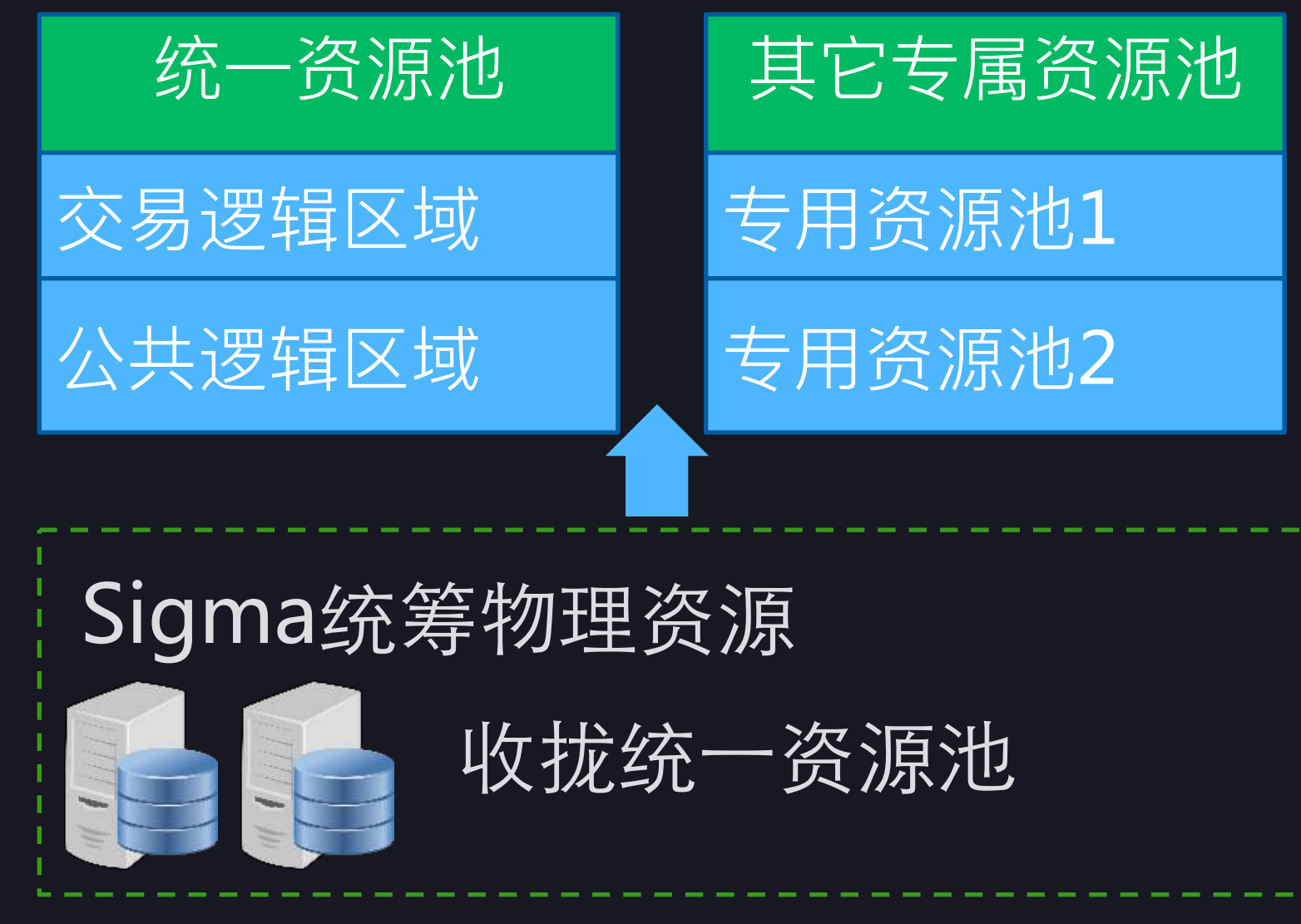
- ✓ **规模：**各T4分组规模不一，大部分都是小规模：资源碎片化。
- ✓ **调度：**T4分组内小规模调度，核心应用打散受限。
- ✓ **资源分配：**双11期间参差不齐：
 - ✧ 交易相关CPU充分售卖，无空闲CPU。
 - ✧ 但众多T4分组，宿主机尚未分配容器实例。
- ✓ **资源利用率非常不均衡：**
 - ✧ 部分分组CPU满负载运行（高达45-50%）
 - ✧ 但相当多的分组CPU几乎完全空闲

系统整体架构和历史演进

●改变：云化架构，混合云

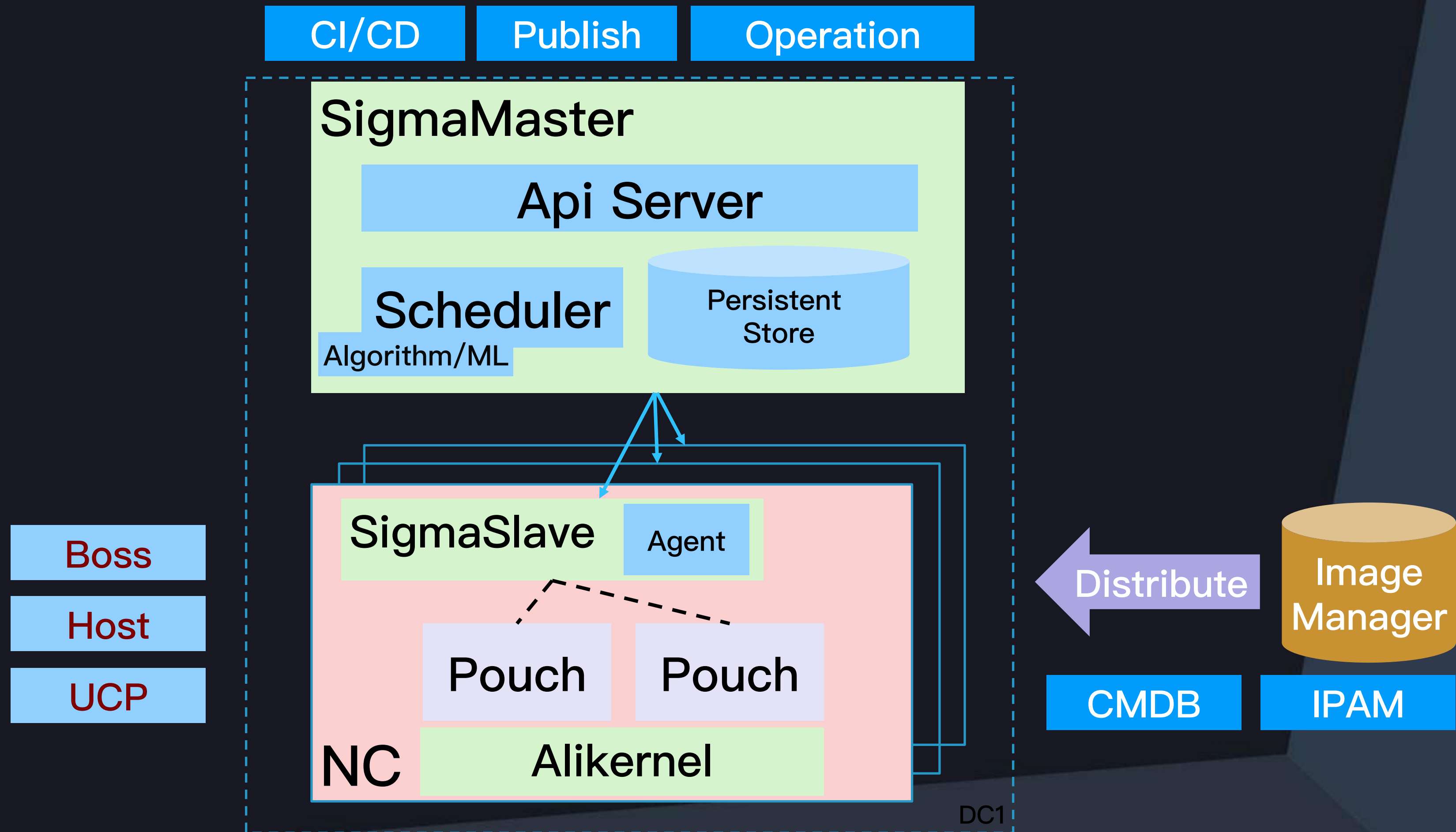
- ✓规模：统一大资源池模式。
- ✓调度：大资源池下，Sigma调度对核心应用的各种策略保障，得以更充分地发挥价值。
- ✓资源分配：双11充分使用了所有资源，没有闲置。
- ✓资源利用率：资源充分均衡使用
- ✓离线和在线任务开始混部

Sigma的双11



Sigma系统整体架构和历史演进

- 兼容Kubernetes API, 和开源社区共建
- 采用阿里Pouch容器 (兼容OCI标准) <https://github.com/alibaba/pouch>



Sigma系统整体架构和历史演进

- 业务架构特点：业务多样化
- 业务场景复杂

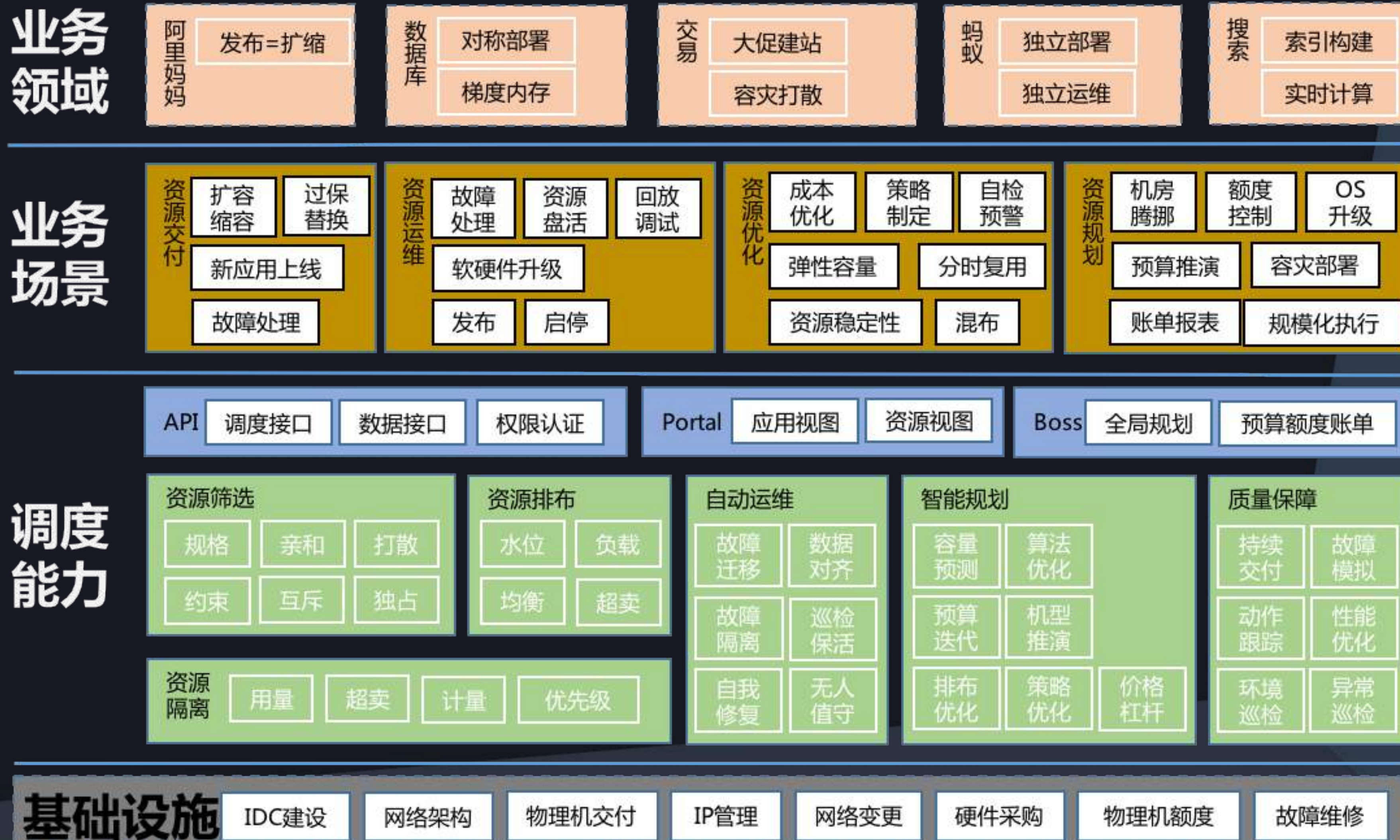


TABLE OF CONTENTS 大纲

- Sigma调度系统整体架构和历史演进
- 架构和实现的特点
- 调度及资源优化
- 展望

Sigma架构和实现特点

- 灵活可配置的调度策略
- 双11给Sigma调度系统带来的挑战以及Sigma的解决方案
 - 复杂约束下的批量调度优化
 - 精确高水位排布
 - 大规模快速建站

灵活可配置的调度策略

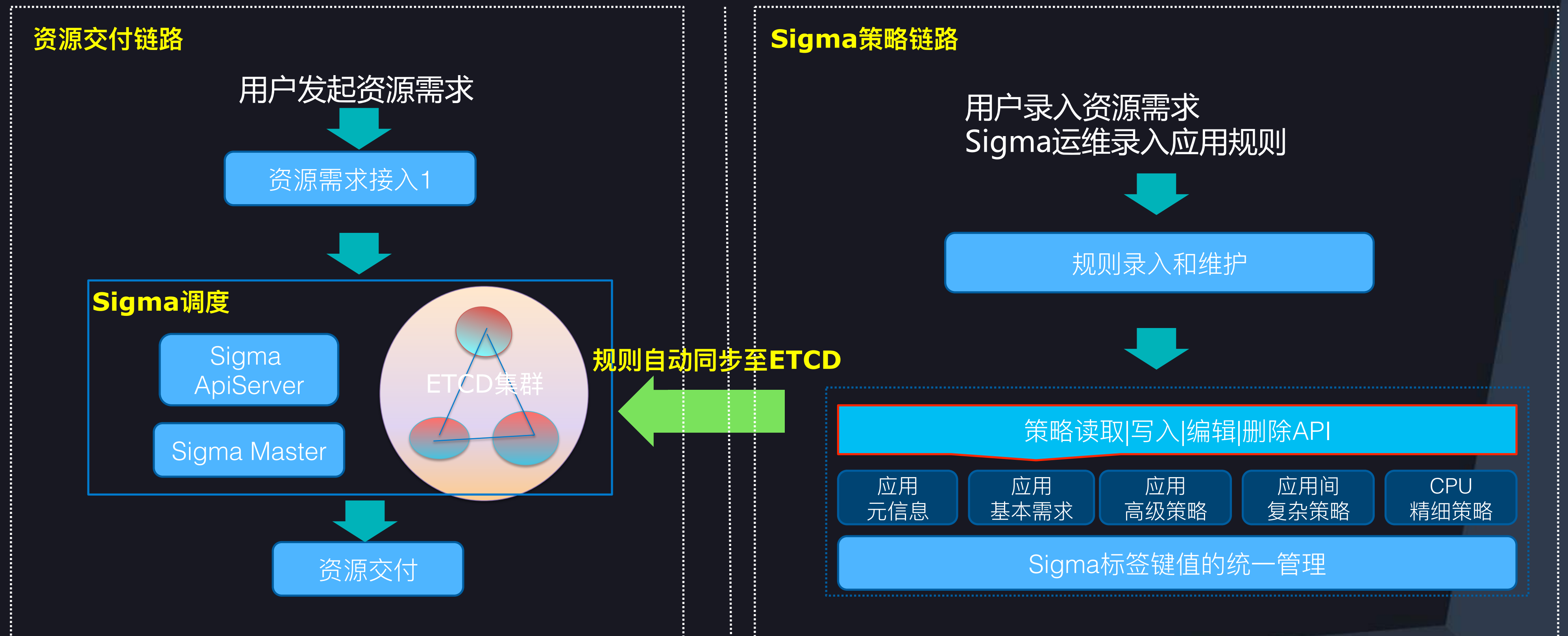
- 支持多样化的应用场景：电商交易、中间件、广告、搜索、数据库、日常测试容器
- 调度优选模型的建立
 - 以插件化的方式基于外部输入实时调控集群打分模型
 - 可配置的优化调度策略，解决资源碎片率高和扎堆严重的问题

灵活可配置的调度策略

- 业务团队开发出新的策略，可立即配置生效，不需要代码发布
- 所有调度策略可配置
- 支持的策略：
 - 应用部署：亲和、互斥、独占、P0M0（最重要优先级应用）
 - 其他策略：资源需求，容器创建特殊需求，Ip隔离需求
 - CPU精细调节：CPUSet独占、均衡、SameCore等策略

Sigma系统整体架构和历史演进

- 调度系统的架构和设计满足业务场景需要



Sigma架构和实现特点

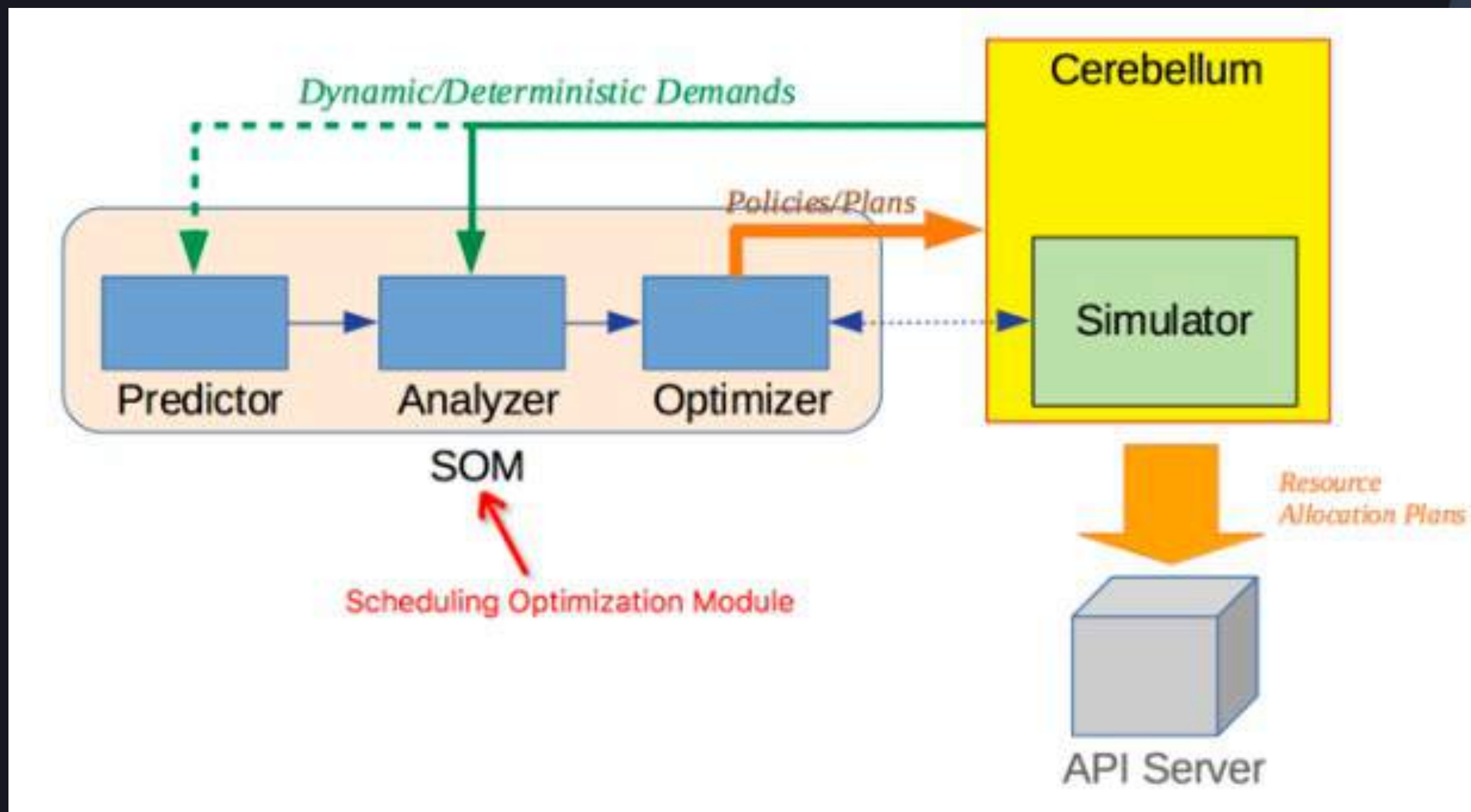
- 灵活可配置的调度策略
- 双11给Sigma调度系统带来的挑战以及Sigma的解决方案
 - 复杂约束下的批量调度优化
 - 精确高水位排布
 - 大规模快速建站
 - 混合云+弹性

双11给调度带来的挑战

- 复杂约束下的批量调度优化
 - 目标：通过批量优化，达到比顺序调度更加优化的效率
 - 实现方式：Scheduling optimization module

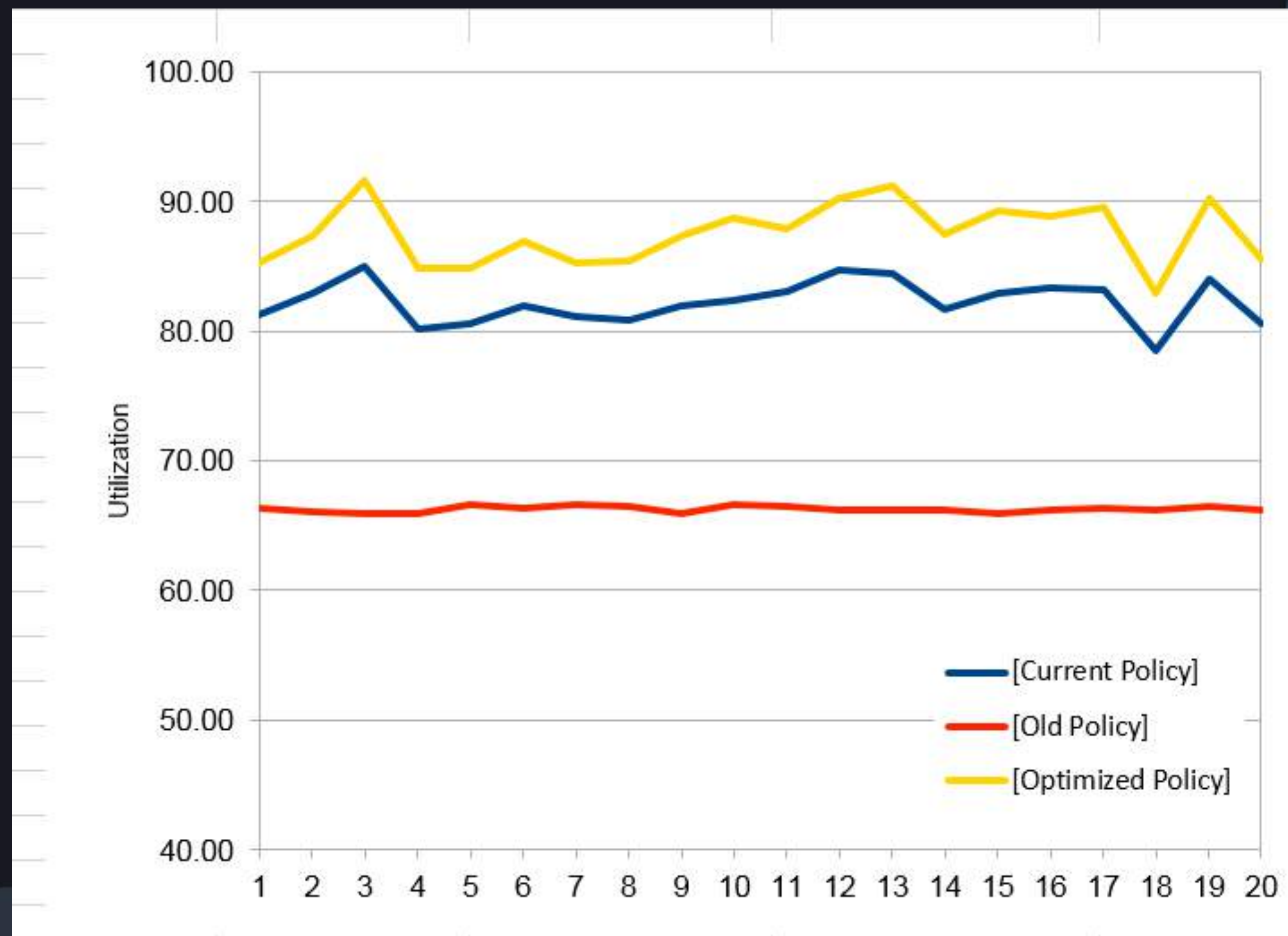
复杂约束下的批量调度优化

- 通过模拟器和线上数据回放，对批量建站请求进行仿真模拟，并优化方案



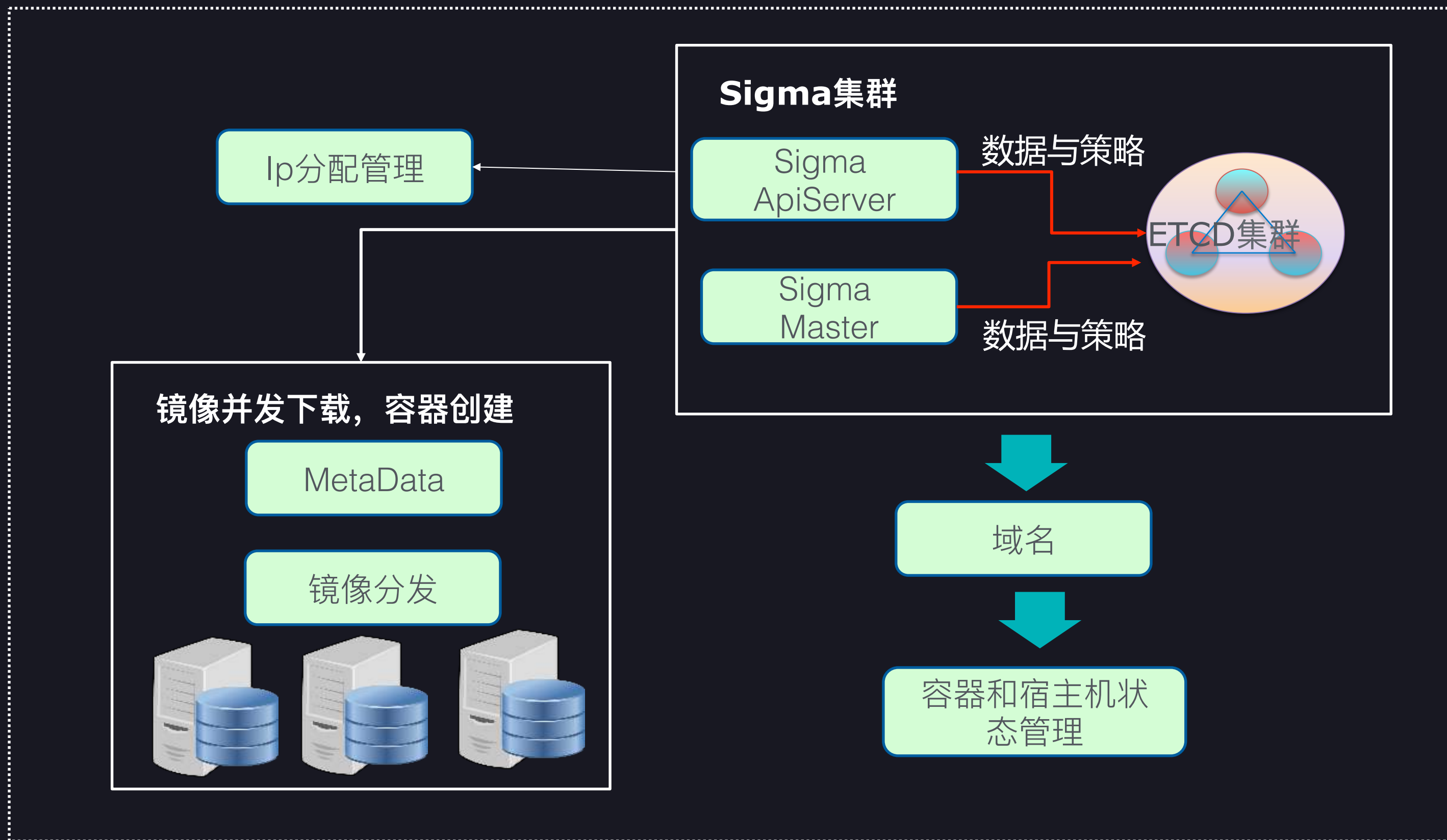
复杂约束下的批量调度优化

- 调度优化效果：分配率从不到70%提升到80%到90%



大规模快速建站

资源交付链路的压测与全链路性能优化



Sigma上下游全链路性能优化：

- Sigma APIServer的并发扩容的优化控制。
- Sigma调度核心master的性能保证。
- ETCD的性能保证和优化。
- 镜像下载：蜻蜓的优化。
- 镜像落盘：调度模拟与宿主机亲和。

大规模快速建站

镜像下载与镜像落盘的优化：

- 日常扩容仅预热基础镜像。
- 大促建站预热基础镜像。
- 大促建站针对性预热最热门的应用镜像。
- 大促建站针对独占资源池：指定具体的应用，提前镜像预热。
- Sigma模拟器：提前计算整单元的部署结果，预热宿主机（基于宿主机亲和的调度规则）。

日常资源交付

基础镜像预热

镜像自动清理工具

下载或P2P分发

镜像分发超级节点集群（按机房或地域部署）



实时分片
任务调度，最优的下载路径

大规模快速建站

大促建站资源交付

建站场景：大规模建站（初始化）

压测场景：变压边弹

建站场景：独占应用部署

模拟器：宿主机应用亲和

最热门应用镜像预热

基础镜像预热

独占资源池：预部署对应独占应用镜像

镜像提前落盘

镜像自动清理工具

下载或P2P分发

镜像分发超级节点集群（按机房或地域部署，提前增加超级节点容量）



实时分片任务调度，最优的下载路径

超级节点提前预热工具

TABLE OF CONTENTS 大纲

- Sigma调度系统整体架构和历史演进
- 架构和实现的特点
- 调度及资源优化
- 展望

调度和资源优化

资源池分散

- 资源碎片化
- 不便于混部

调度排布效率

- 排布效率不高, Stranded resources
- 排布效率: 复用

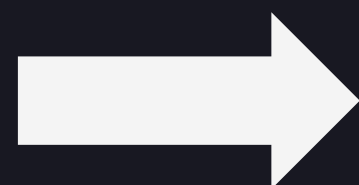
资源规划和预约

- 为峰值需求保留资源
- 不确定性成本

调度和资源优化

资源池分散

- 资源碎片化
- 不便于混部和复用

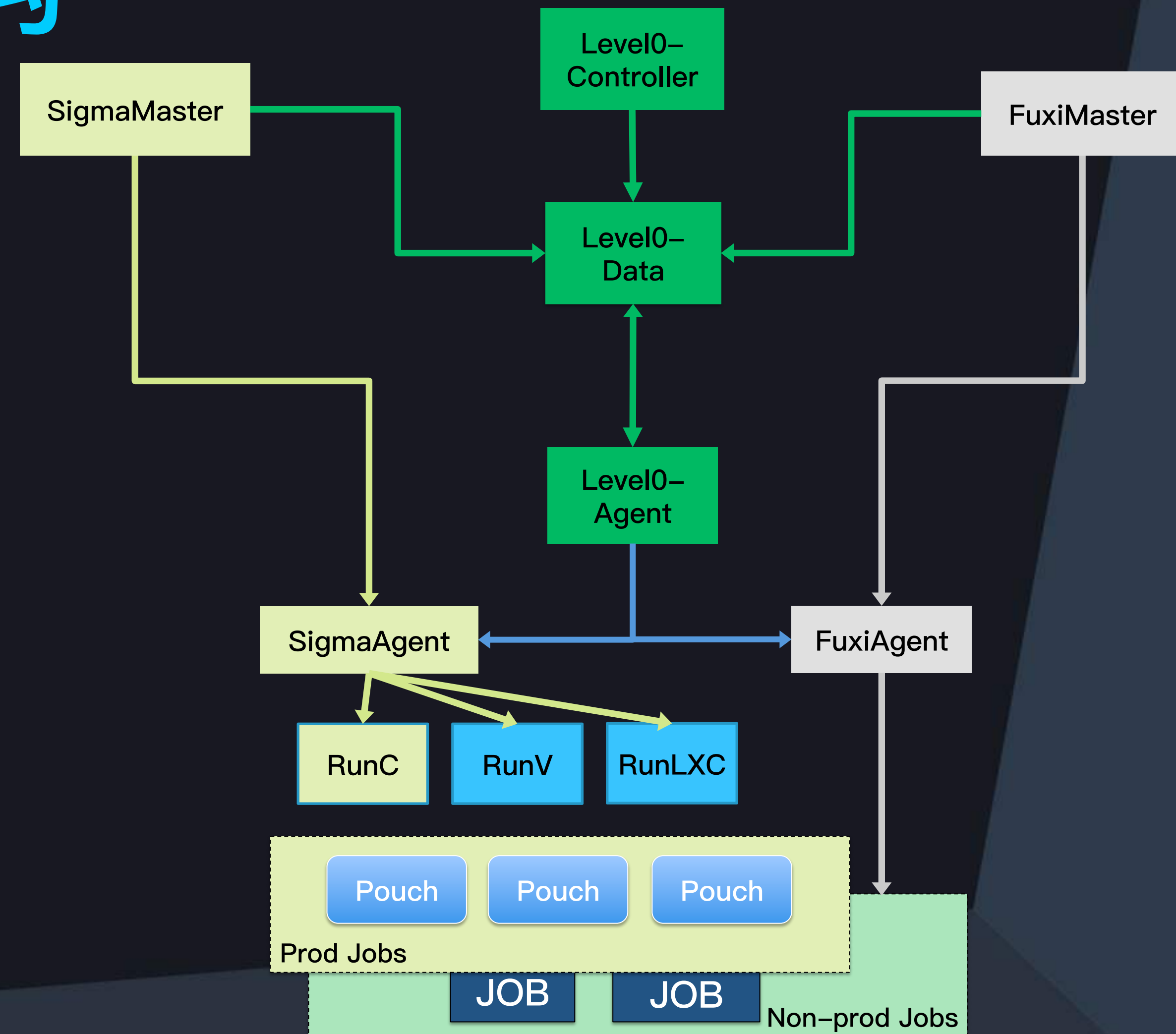


统一资源池和底层调度

在线离线任务混部

Sigma与Fuxi混布架构

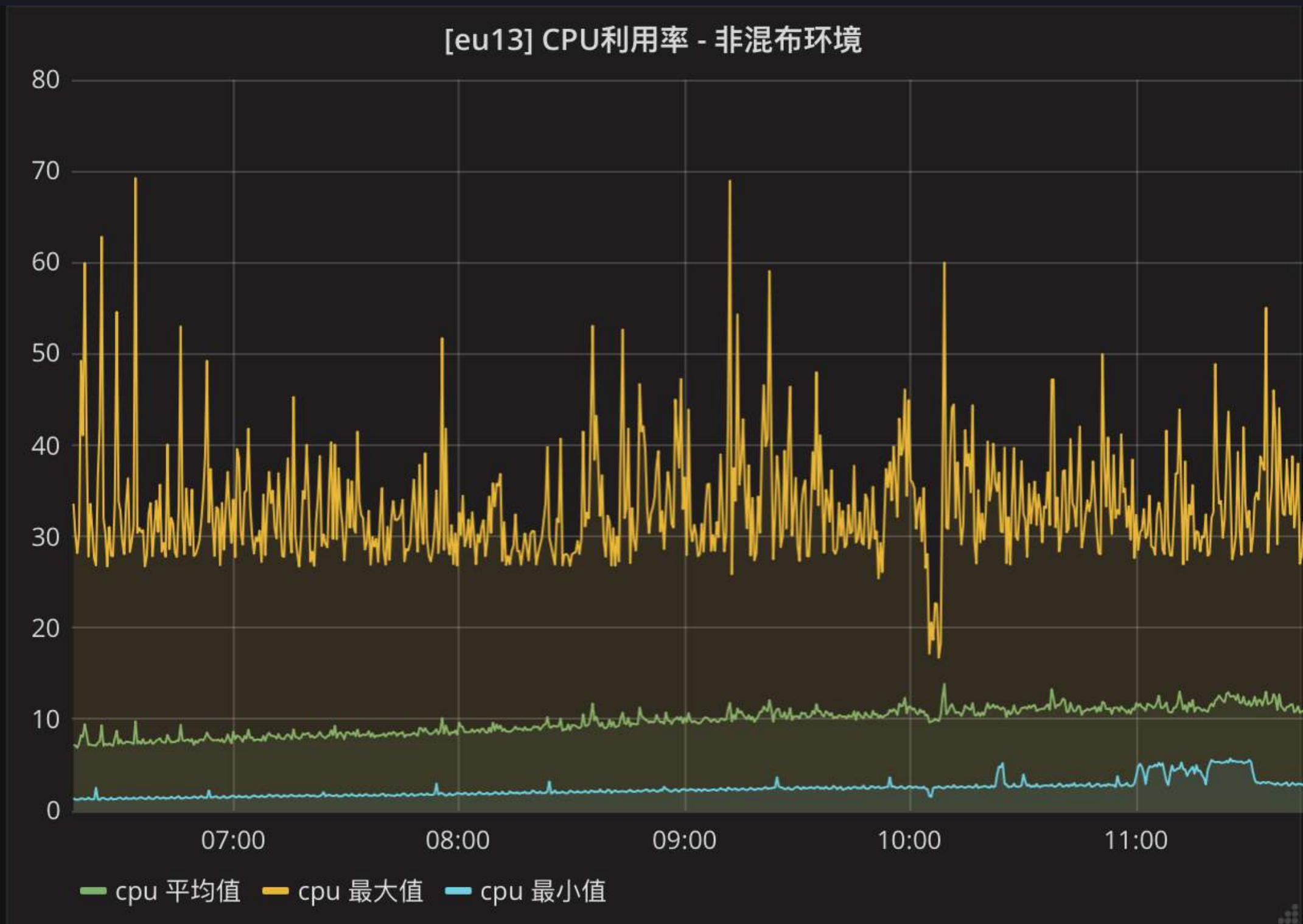
- 通过Sigma和Fuxi完成在线、离线各自的调度，离线共享超卖
- 通过零层相互协调资源配比做混布决策，通过内核解决资源竞争隔离的问题
- 属于共享状态(shared state)的分布式调度 <http://www.firmament.io/blog/scheduler-architectures.html>



混部关键技术

- 内核资源隔离
 - CPU HT资源隔离：Noise Clean内核特性，解决在、离线超线程资源争抢问题
 - CPU 调度隔离：CFS基础上增加Task Preempt特性，提高在线任务调度优先级
 - CPU 缓存隔离：CAT，在、离线三级缓存(LLC)通道隔离(Broadwell及以上)
 - 内存隔离：CGroup隔离/OOM优先级；Bandwidth Control减少离线配额实现带宽隔离
 - 网络QoS隔离：管控打标为金牌；在线打标为银牌；离线打标为铜牌，分级保障带宽

混部对利用率提升



- 通过混部，CPU日均利用率从10%提升到40+%

调度和资源优化

- List 1

- List 2

调度排布效率

- 排布效率不高, Stranded resources
- 排布效率: 复用

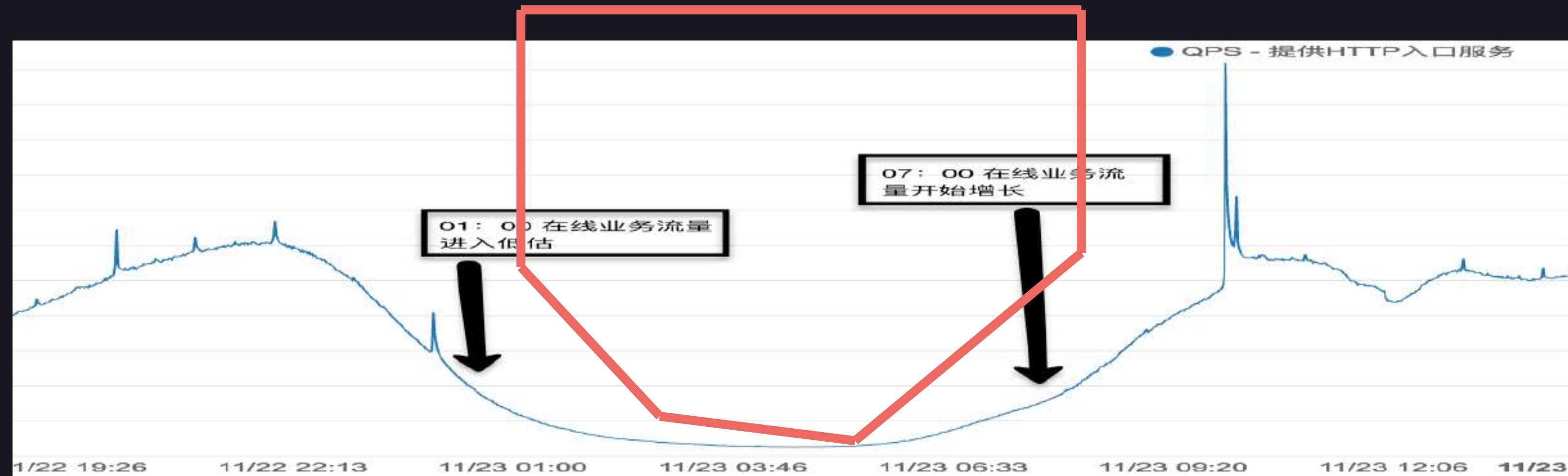


在线任务优化排布

分时复用

弹性调度和资源分时复用

全天的流量走势



离线扩容
在线缩容

离线缩容
在线扩容

- 结合分时复用，利用率可以进一步提升
- 挑战：
 - 容量预测
 - 扩缩容稳定性
 - SLO保障

TABLE OF CONTENTS 大纲

- Sigma调度系统整体架构和历史演进
- 架构和实现的特点
- 调度及资源优化
- 展望

展望

- 资源最优
 - 混合云架构进一步发展，混合云深度联动，实时共享资源
 - 通过大规模混部、优先级差异化提升资源使用效率
- 智能化调度：更好的感知应用的SLA需求、减少应用间干扰
- 适应异构计算需求，为异构计算优化

感兴趣继续讨论？
加入我们一起挑战？

THANK YOU

如有需求，欢迎至 [讲师交流会议室] 与我们的讲师进一步



archsummit 阿里云化专场



该二维码7天内(12月13日前)有效，重新进入
将更新

A promotional graphic for the ArchSummit Alibaba Cloud Special Event. It features a small grid of images in the top left, the event title 'archsummit 阿里云化专场', a large central QR code, and a note at the bottom stating the QR code is valid for 7 days (before Dec 13) and will be updated upon re-entry.