ArchSummit
全球架构师峰会 2017

主办方 Geekbang 极客邦科技 InfoQ

腾讯云大规模任务调度系统的架构蜕变

王旻
腾讯云 高级技术专家

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

主办方 Geekbang. 极客邦科技 InfoQ

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

王旻 alexmwang

腾讯云 高级技术专家

硕士就读于中科院计算所，有丰富的分布式调度系统理论和实践经验。

2015年加入腾讯，负责腾讯云CVM（云主机）和批量计算产品的设计和开发，致力于打造高吞吐、高可用的调度系统和计算产品。

TABLE OF

CONTENTS 大纲

- 聚焦任务调度
- 任务调度的核心挑战
- 调度系统架构蜕变
- 调度系统实现细节
- 总结与心得

TABLE OF

CONTENTS 大纲

- 聚焦任务调度
- 任务调度的核心挑战
- 调度系统架构蜕变规律
- 调度系统设计与实现
- 心得体会

聚焦任务调度

- 分布式调度
  - 用户之间，优先为哪些用户分配资源
  - 任务之间，优先为哪些任务分配资源
  - 任务调度，为任务分配机器资源
- Google[1]：task scheduling refers to the assignment of tasks to machines.

聚焦任务调度

- 任务调度
  - Task 和 Machine
- 公有云中的任务调度
  - VM 和 HOST
  - Task -> VM
  - Machine -> Host
  - 任务调度：为 VM 分配 HOST

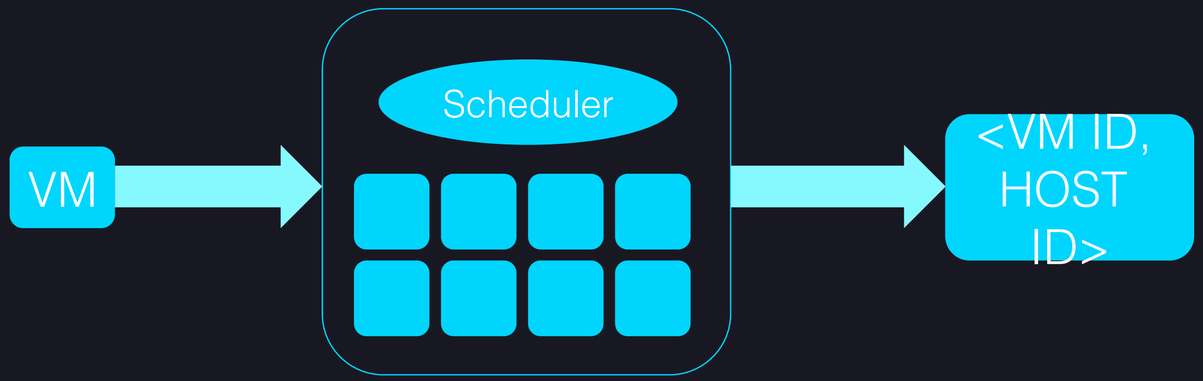


TABLE OF

CONTENTS 大纲

- 聚焦任务调度
- 任务调度的核心挑战
- 调度系统架构蜕变规律
- 调度系统设计与实现
- 心得体会

异构性 与 调度质量

- HOST
  - 数据中心维护周期长、集群规模大，不同批次 HOST 在软硬件存在不同
  - HOST 新特性灰度
- VM
  - VM 不同实例机型，例如不同代次，GPU、FPGA 等
  - VM 反亲和性，并发创建打散、镜像缓存
- 趋势
  - 不是所有 HOST 都能满足 VM 的需求 —— 硬性约束
  - 满足 VM 需求的 HOST，其满足程度是不同的 —— 软性约束
  - VM 和 HOST 是调度的主角，异构性增加了二者匹配的复杂度，必须考虑约束

可扩展性 与 调度吞吐率

- HOST
  - 单Region，数万台 物理服务器
- VM
  - 云计算需求爆发式增长，潮汐式海量并发购买
    - CVM 直接用户：爬虫、秒杀抢购
    - CVM 间接用户：弹性伸缩、批量计算、竞价实例
    - 规模大，时效性强
  - 每小时 数万台 VM 购买请求，峰值每分钟 上千台 VM 购买请求
- 问题
  - CVM 当时的生产吞吐率为 100台/分钟，无法满足用户海量购买请求
  - Scheduler 成为整个系统的性能瓶颈，调度吞吐率不足，处理延迟增加，影响系统的可扩展性
  - 用户等待时间延长
  - 同时影响业务时效性和用户体验

核心挑战

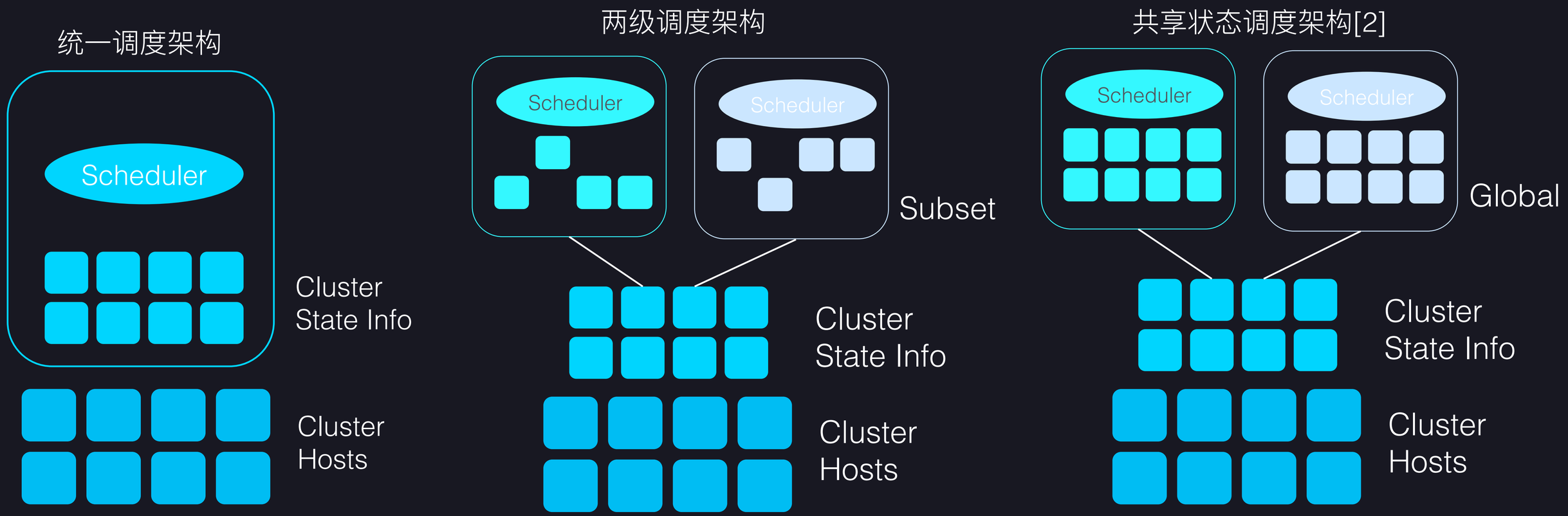
- 保证调度质量的前提下，显著提升调度吞吐率

TABLE OF

CONTENTS 大纲

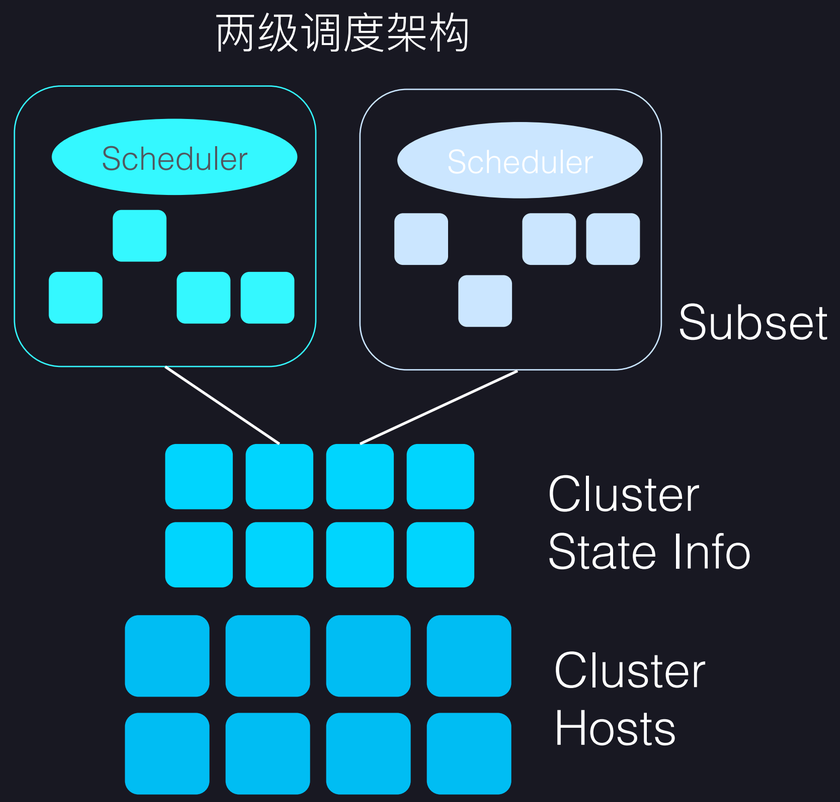
- 聚焦任务调度
- 任务调度的核心挑战
- 调度系统架构蜕变规律
- 调度系统设计与实现
- 总结与心得

调度系统架构演变

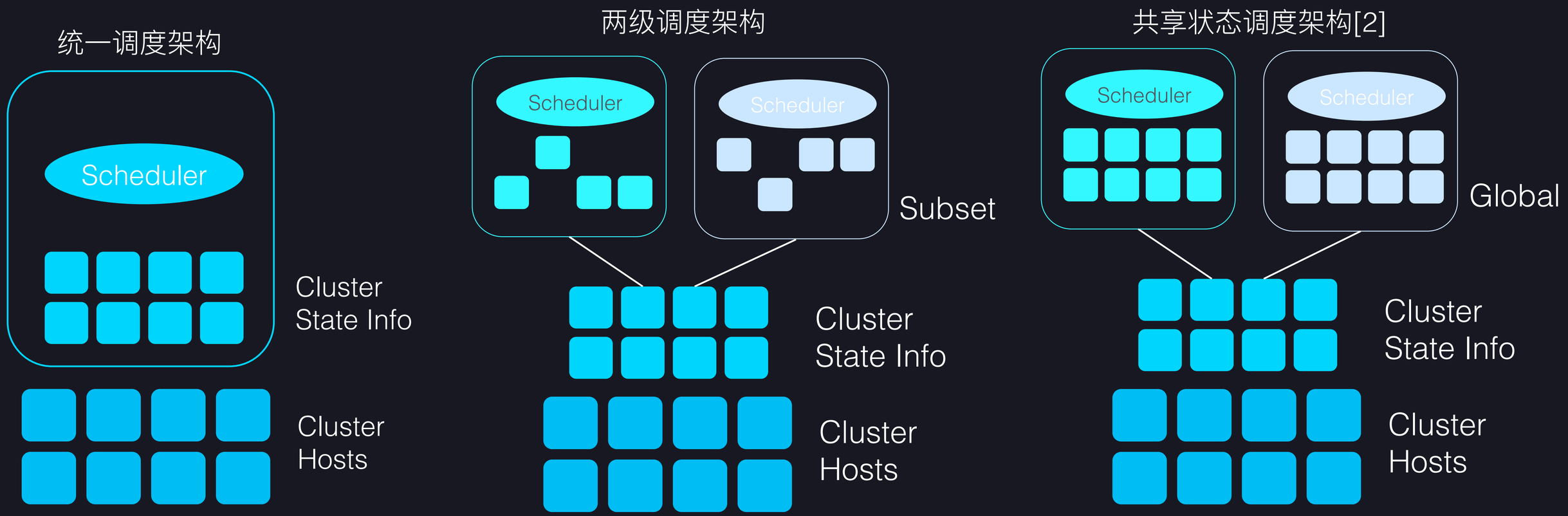

统一调度架构
Scheduler
Cluster
State Info
Cluster
Hosts
两级调度架构
Scheduler
Scheduler
Subset
Cluster
State Info
Cluster
Hosts
共享状态调度架构[2]
Scheduler
Scheduler
Global
Cluster
State Info
Cluster
Hosts

调度系统架构演变

- 典型代表
  - Mesos，通过 Resource Offer 和上层调度器通信
  - 实现多个 Framework 共享集群资源
- 局限
  - 无全局资源视图
    - 无法保证调度决策全局最优，无法跨调度器抢占
  - 并发度
    - Resource Offer 本质上是在不同 Framework 之中串行轮询，并发度仍有提升空间。



调度系统架构演变


统一调度架构
Scheduler
Cluster
State Info
Cluster
Hosts
两级调度架构
Scheduler
Scheduler
Subset
Cluster
State Info
Cluster
Hosts
共享状态调度架构[2]
Scheduler
Scheduler
Global
Cluster
State Info
Cluster
Hosts

调度系统架构演变

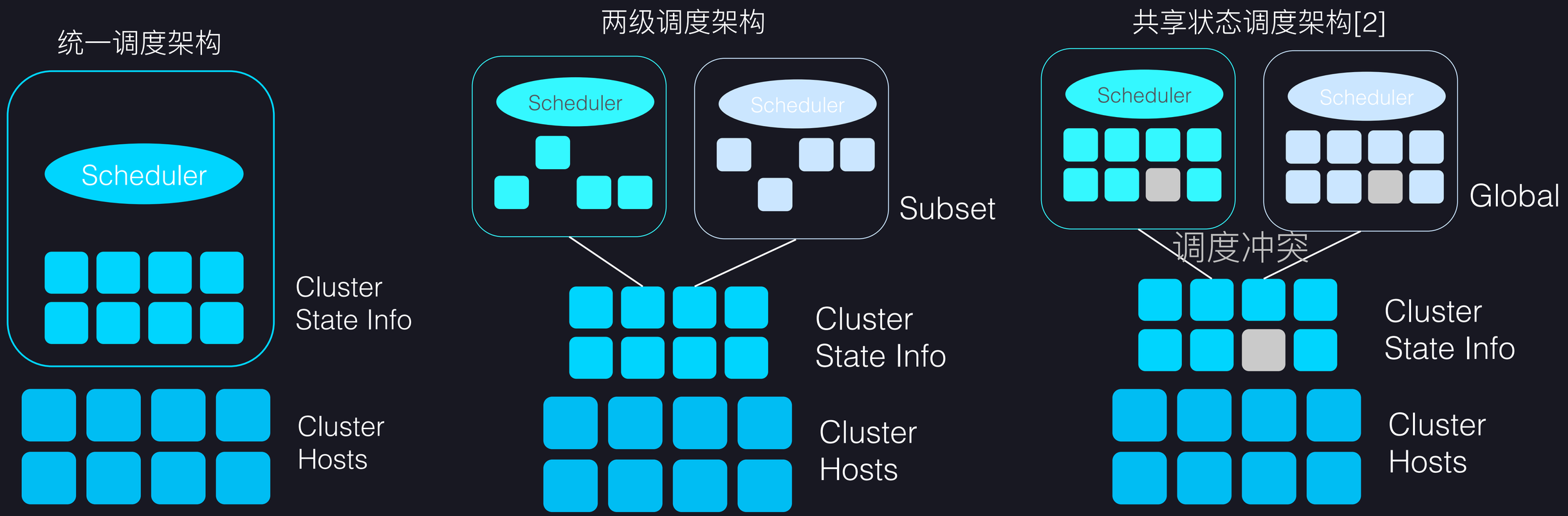

统一调度架构
Scheduler
Cluster
State Info
Cluster
Hosts
两级调度架构
Scheduler
Scheduler
Subset
Cluster
State Info
Cluster
Hosts
共享状态调度架构[2]
Scheduler
Scheduler
Global
调度冲突
Cluster
State Info
Cluster
Hosts

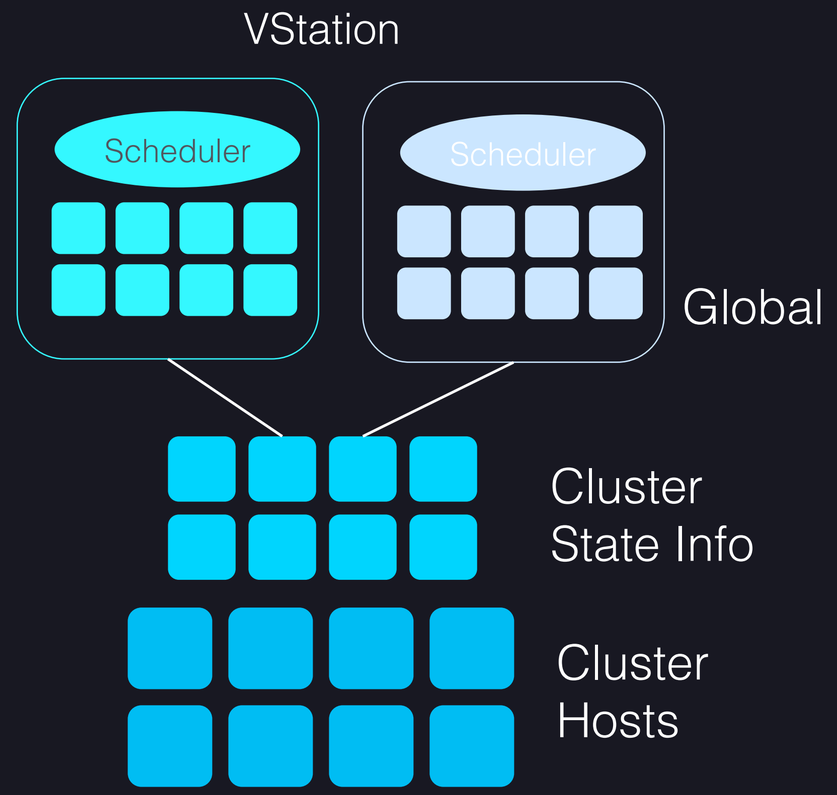
TABLE OF

CONTENTS 大纲

- 聚焦任务调度
- 任务调度的核心挑战
- 调度系统架构蜕变规律
- 调度系统设计与实现
- 心得体会

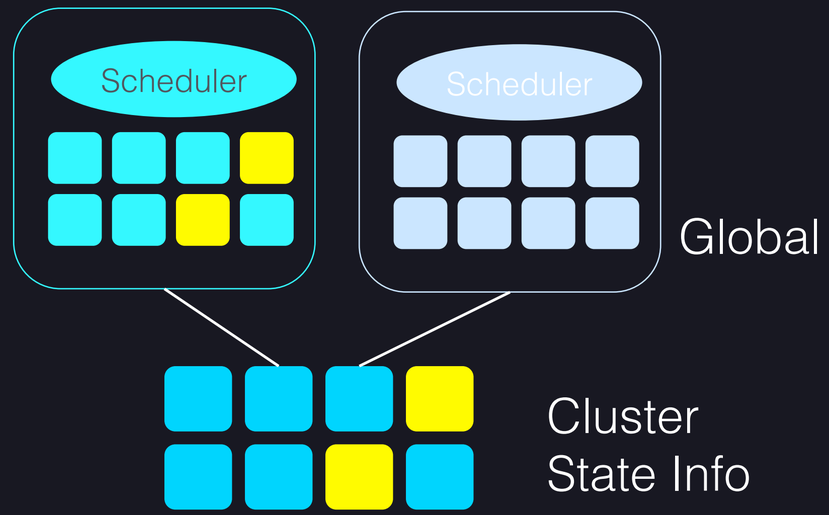
CVM VStation

- 共享状态调度架构
  - 多调度器，并发调度
  - 基于全局资源视图，支持调度算法最优解
  - 乐观无锁并发，提交调度结果保证事务性
  - 优化调度冲突
- 调度流程
  - 资源同步
  - 调度决策
  - 提交结果



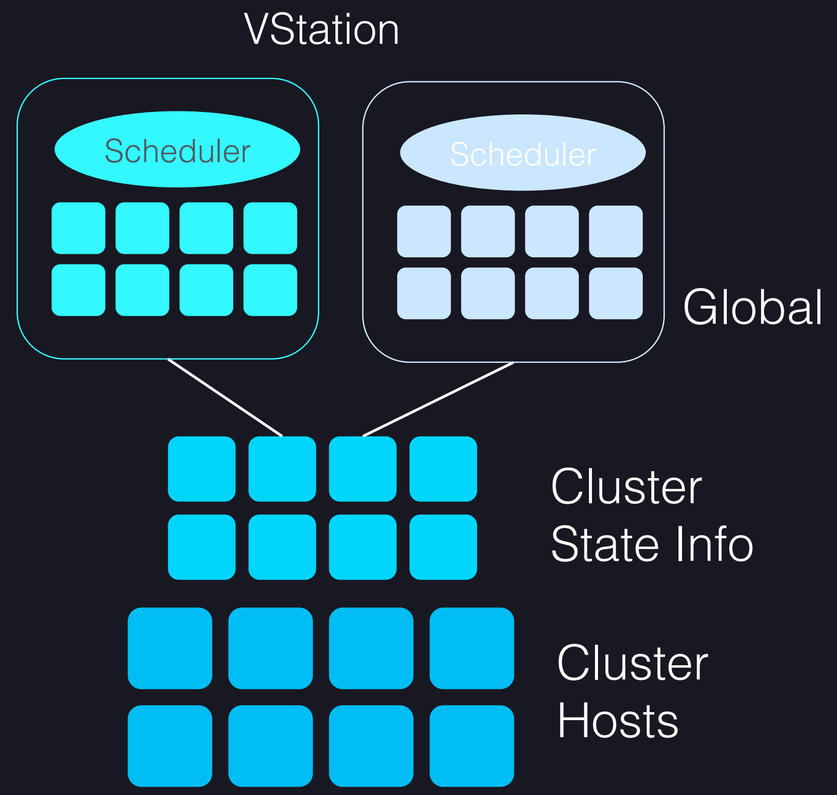
CVM VStation 实现细节

- 资源同步
  - 调度器拉取集群状态信息
  - HOST 数万规模，调度器数百规模
  - 调度器私有缓存 + 增量更新
    - 首次启动全量更新
    - 后续增量更新
  - 同步数据量平均减少90%以上



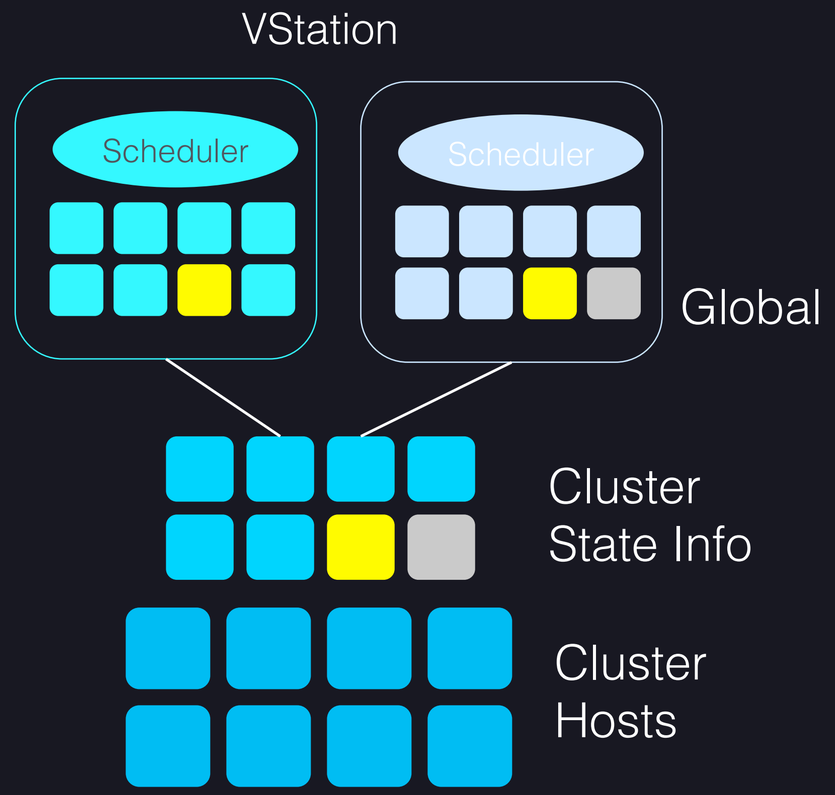
CVM VStation 实现细节

- 调度决策
  - 过滤
    - 排除不符合硬性约束的 HOST
  - 排序
    - 根据反亲和性、镜像缓存、资源利用率等维度，进行多维排序
  - 随机打散
    - 对于前K个HOST进行随机打散，防止调度冲突



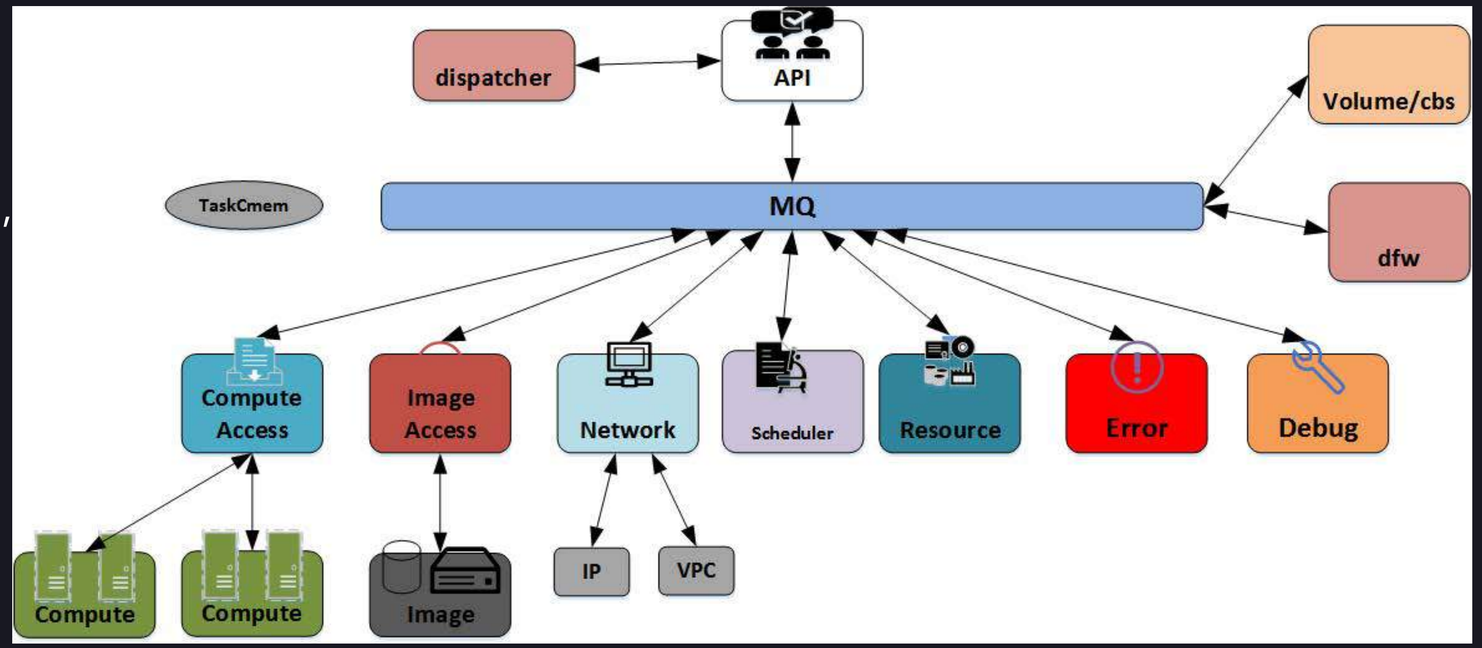
CVM VStation 实现细节

- 提交结果
  - 按序遍历 HOST 候选列表，模拟扣减资源
  - 提交资源变更事务：资源数据、反亲和性记录
  - 事务成功
    - 则调度成功，同时更新私有缓存中的数据
  - 事务多次失败
    - 发生调度冲突，尝试下一台 HOST



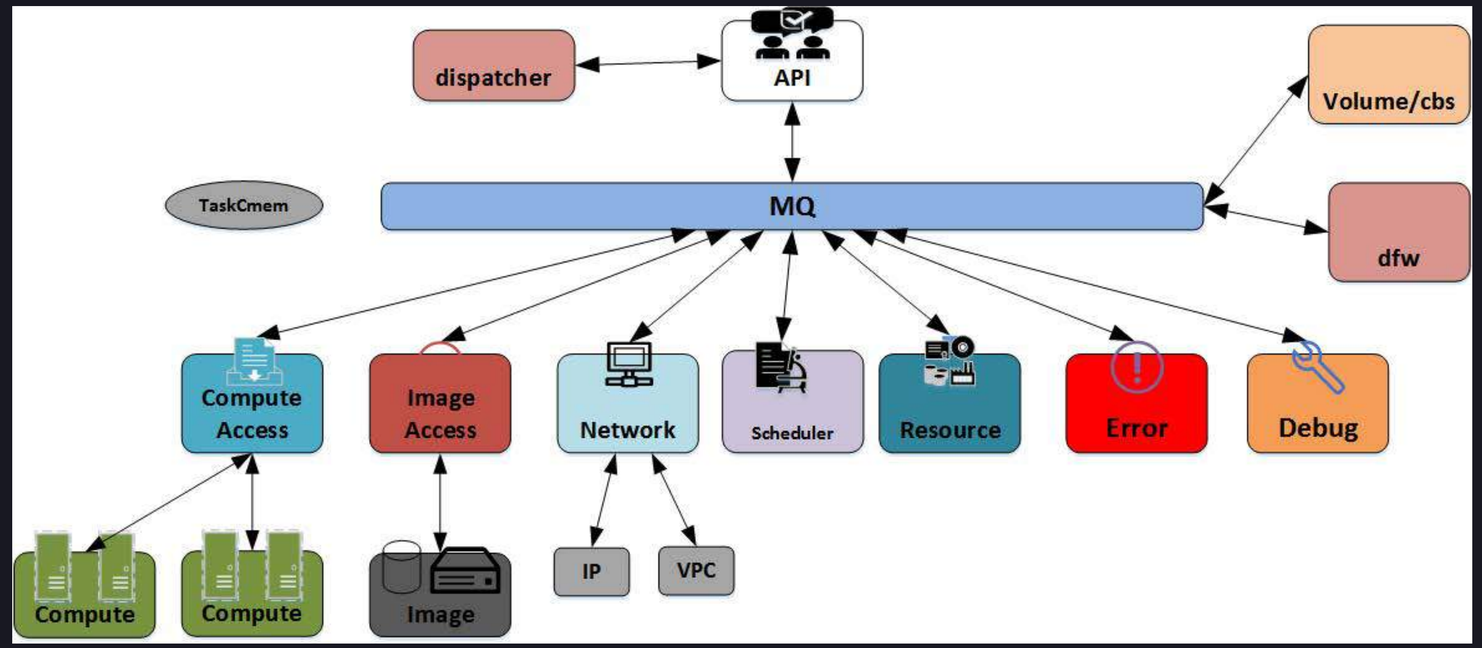
CVM VStation 其他细节与优化

- 消息流转
  - 通过MQ
  - 通过step_config配置流程步骤
  - 根据每一步骤的配置投递到指定消息队列，再由消费者进行处理
- 内部回滚
  - 某个步骤失败后，根据step_config配置生成回滚流程，开始回滚，保证流程事务性



CVM VStation 其他细节与优化

- 消息压缩
  - 兼顾数据压缩比、压缩速率、以及对于资源的额外开销
  - 压缩比 20%



CVM VStation 其他细节与优化

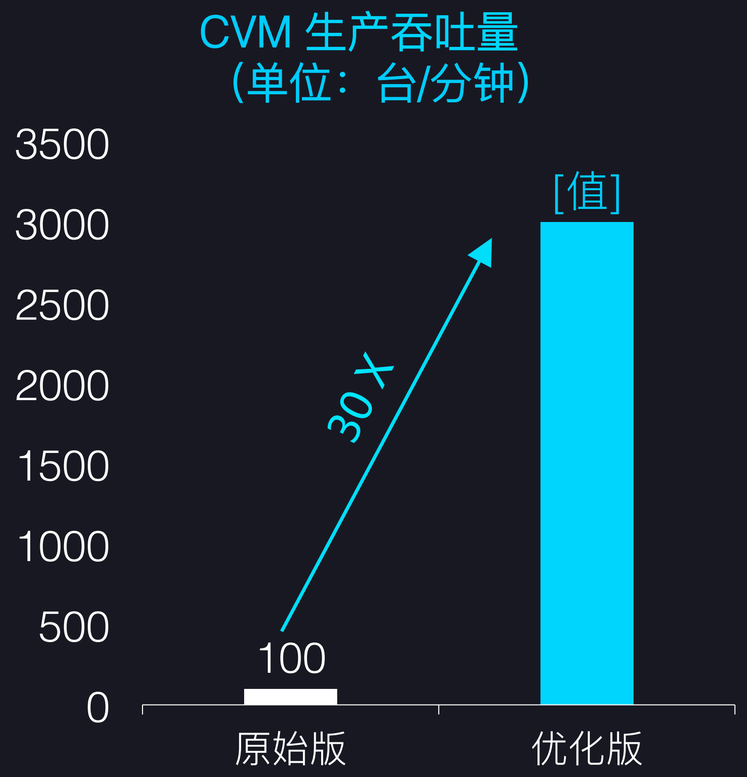
- 镜像缓存
  - HOST 缓存高频使用镜像，主要是公有镜像
  - 调度策略，同等条件下，优先选择命中镜像缓存的 HOST

CVM VStation 其他细节与优化

- CBS快照回滚
  - CBS 是腾讯云云盘产品简称
  - 创建使用云盘的CVM，如果CBS后台存在对应快照，会进行秒级快照回滚，避免下载镜像，显著减少创建时间

应对海量并发创建

- HOST 数万台
- Scheduler 数百个
- 生产吞吐率 提升30倍
- 生产时间 下降90%



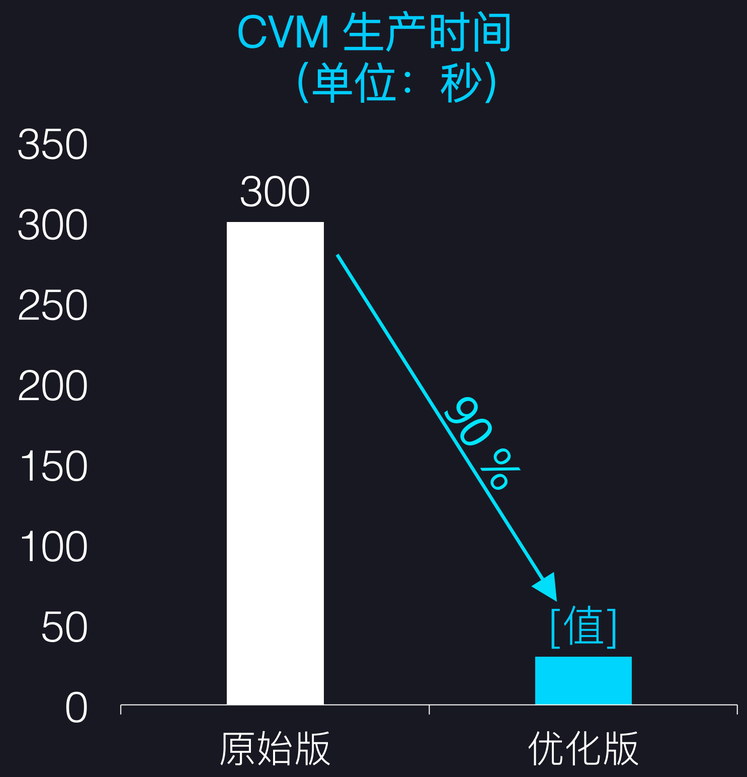


TABLE OF

CONTENTS 大纲

- 聚焦任务调度
- 任务调度的核心挑战
- 调度系统架构蜕变规律
- 调度系统设计与实现
- 心得体会

心得体会

- 系统演化
  - 面对相同挑战，不同系统可能会进化成为相近的样子
  - 不同文明都独自发明出轮子
- 异构化是客观挑战，而非主观追求
  - 异构化造成资源的逻辑分化，与云计算初衷相对立
  - 需求方提供明确的灰度和资源供给计划，防止资源不足

总结与心得

- 评价标准
  - 调度处于系统中央
  - 需求方立场不同，对调度器的要求和评价标准也不同

交流沟通

- 技术交流
  - 微信：wangmin583865
- 求贤若渴
  - alexmwang@tencent.com
  - 北京、深圳
  - 调度，云主机，弹性伸缩，批量计算

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交流

ArchSummit
全球架构师峰会 2017

异构性 与 调度质量

- 调度策略：过滤 + 排序
- 过滤
  - 硬性约束：必须满足的条件
  - 排除不符合条件的
- 排序
  - 软性约束：优先满足的条件
  - 对候选 HOST 进行多维度优先级排序