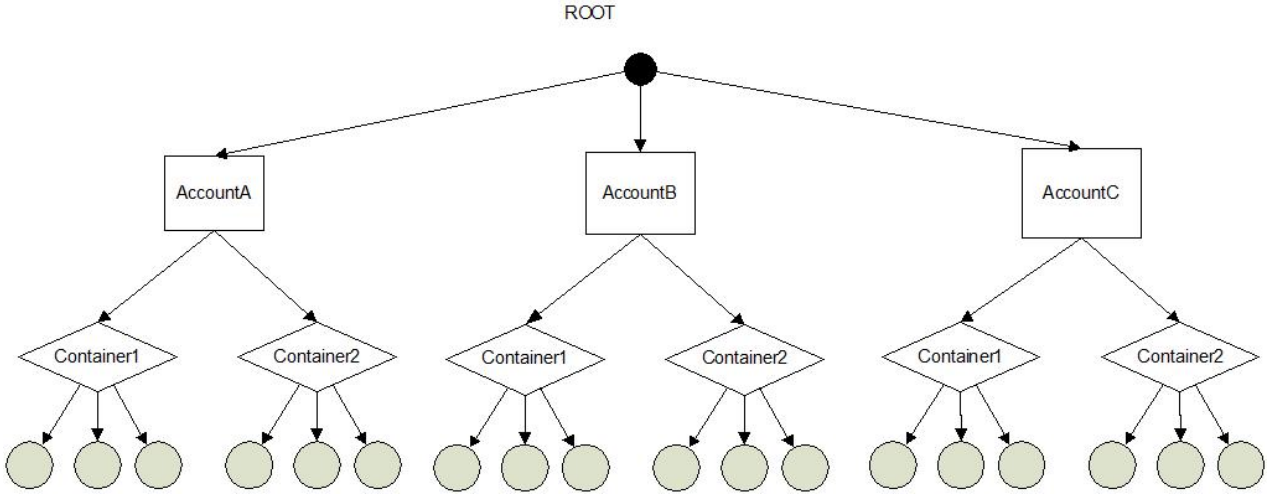
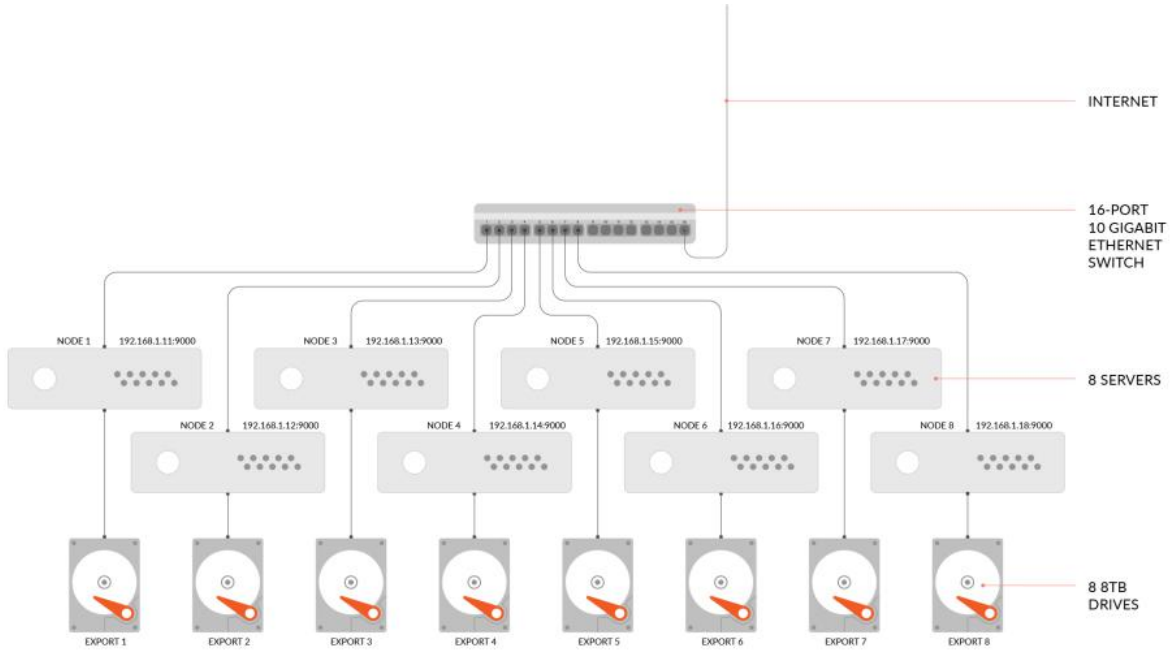
第二部分

尾延迟预测

1. Understanding the latency distribution of cloud object storage systems[J]. JPDC 2019
2. Predicting Response Latency Percentiles for Cloud Object Storage Systems[C]. ICPP 2017

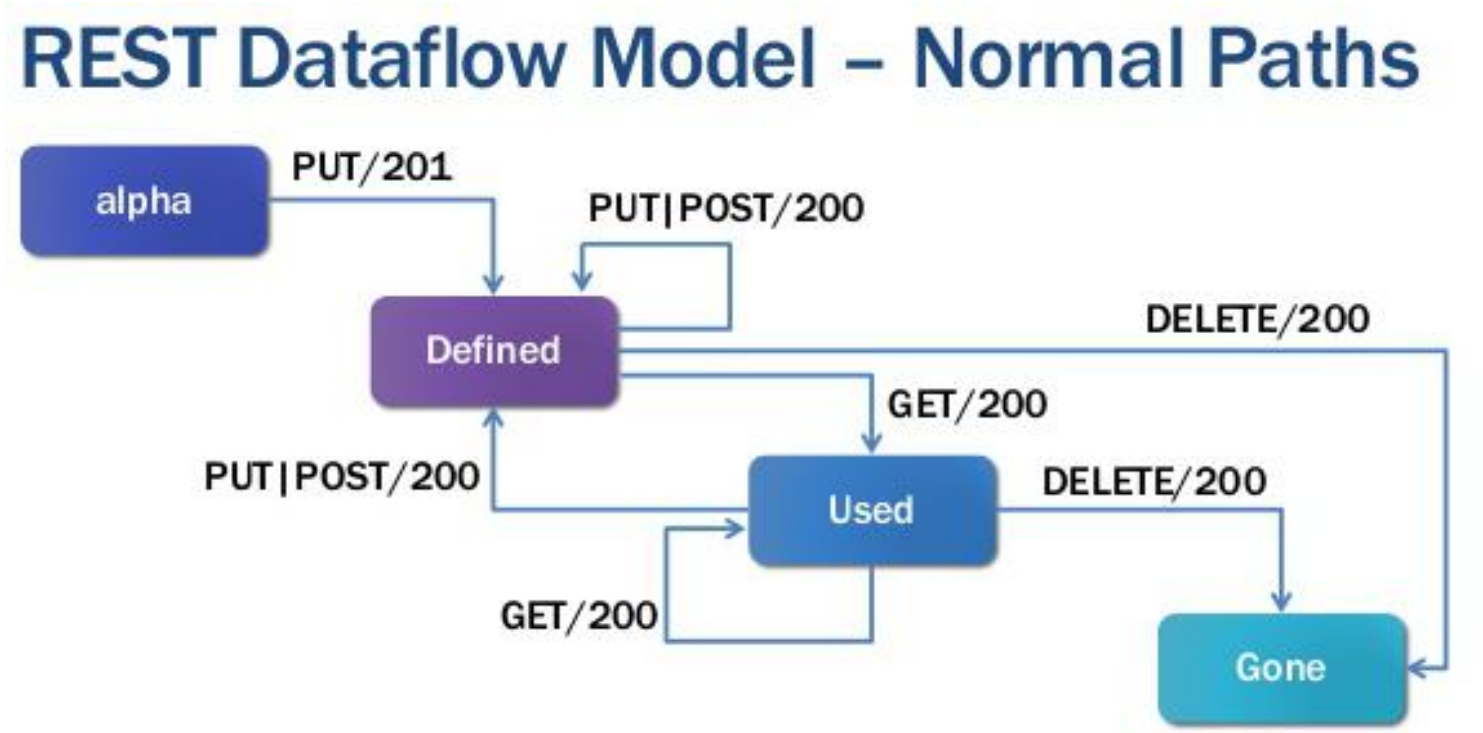
重温尾延迟问题


ROOT
AccountA
AccountB
AccountC
Container1
Container2
Container1
Container2
Container1
Container2


INTERNET
16-PORT
10 GIGABIT
ETHERNET
SWITCH
8 SERVERS
8 8TB
DRIVES
NODE 1
192.168.1.11:9000
NODE 2
192.168.1.12:9000
NODE 3
192.168.1.13:9000
NODE 4
192.168.1.14:9000
NODE 5
192.168.1.15:9000
NODE 6
192.168.1.16:9000
NODE 7
192.168.1.17:9000
NODE 8
192.168.1.18:9000
EXPORT 1
EXPORT 2
EXPORT 3
EXPORT 4
EXPORT 5
EXPORT 6
EXPORT 7
EXPORT 8

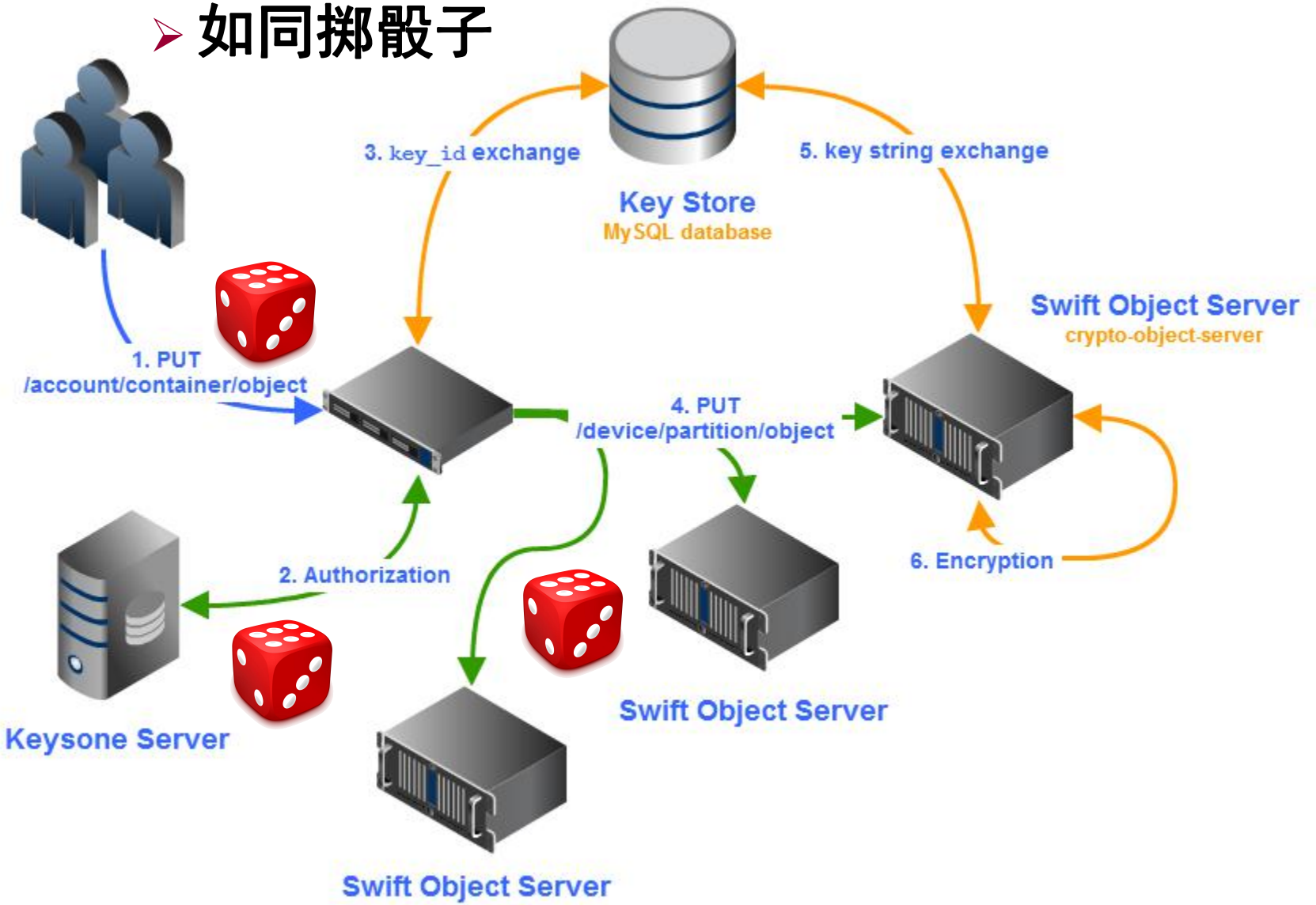
请思考

- 尾延迟挥之不去，问题出在哪里？
  - I/O过程存在随机性



请思考

- 如同掷骰子


3. key_id exchange
Key Store
MySQL database
5. key string exchange
Swift Object Server
crypto-object-server
1. PUT
/account/container/object
4. PUT
/device/partition/object
6. Encryption
2. Authorization
Keysone Server
Swift Object Server
Swift Object Server

做一下数学

Probability of getting a head in a single toss of a coin $p = \frac{1}{2}$

Probability of not getting a head in a single toss of a coin $q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$

Number of coins tossed $n = 6$

4 or more heads means $X \geq 4$ (4, 5, 6)

$$P(X \geq 4) = {}^6C_4 p^4 q^{6-4} \times {}^6C_5 p^5 q^{6-5} \times {}^6C_6 p^6 q^{6-6}$$

$$= {}^6C_4 p^4 q^2 \times {}^6C_5 p^5 q^1 \times {}^6C_6 p^6 q^0$$

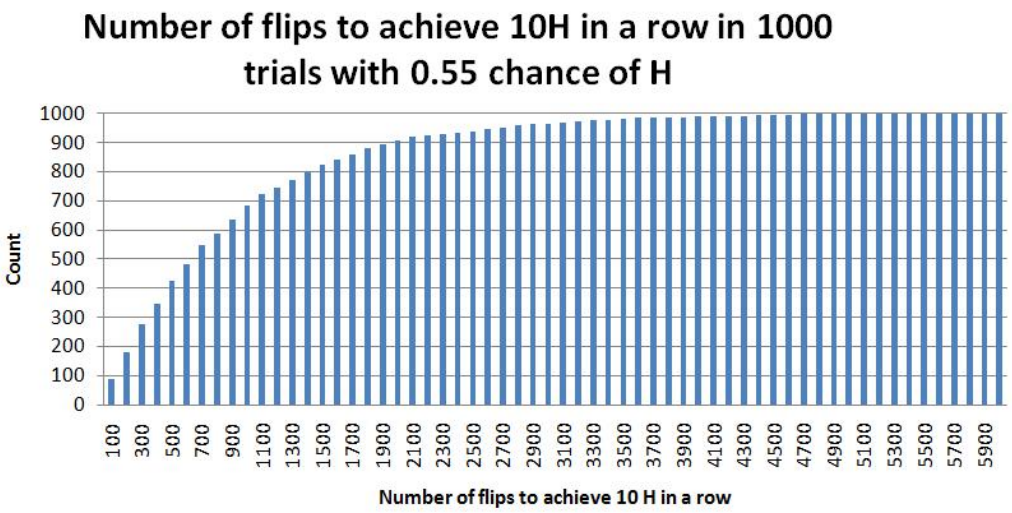
$$= \frac{6!}{4! \times 2!} \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^2 + \frac{6!}{5! \times 1!} \times \left(\frac{1}{2}\right)^5 \times \left(\frac{1}{2}\right)^1 + \frac{6!}{6! \times 0!} \times \left(\frac{1}{2}\right)^6 \times \left(\frac{1}{2}\right)^0$$

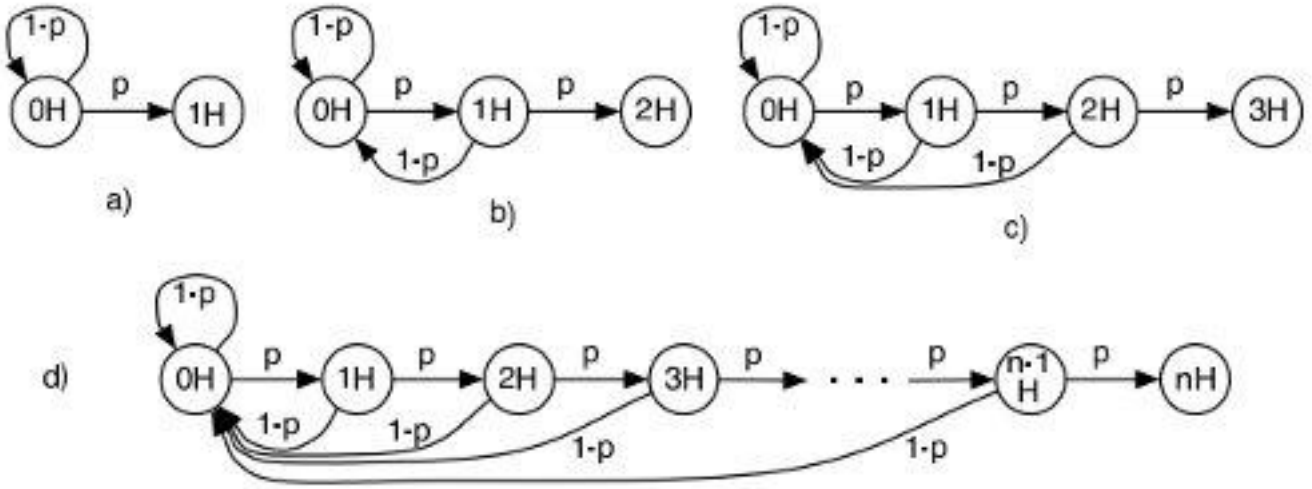
$$= \frac{6!}{4! \times 2!} \times \left(\frac{1}{2}\right)^6 + \frac{6!}{5! \times 1!} \times \left(\frac{1}{2}\right)^6 + \frac{6!}{6! \times 0!} \times \left(\frac{1}{2}\right)^6$$

$$= \left(\frac{6!}{4! \times 2!} + \frac{6!}{5! \times 1!} + \frac{6!}{6! \times 0!}\right) \times \left(\frac{1}{2}\right)^6$$

$$= \left(\frac{6 \times 5 \times 4!}{4! \times 2 \times 1} + \frac{6 \times 5!}{5! \times 1} + \frac{6!}{6! \times 1}\right) \times \frac{1}{64}$$

$$= (15 + 6 + 1) \times \frac{1}{64} = \frac{22}{64} = \frac{11}{32}$$



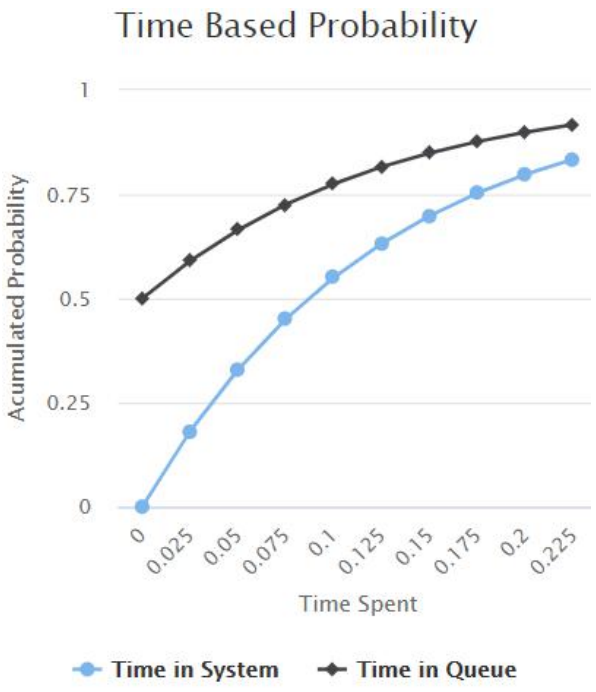


http://www.gwydir.demon.co.uk/jo/probability/info.htm

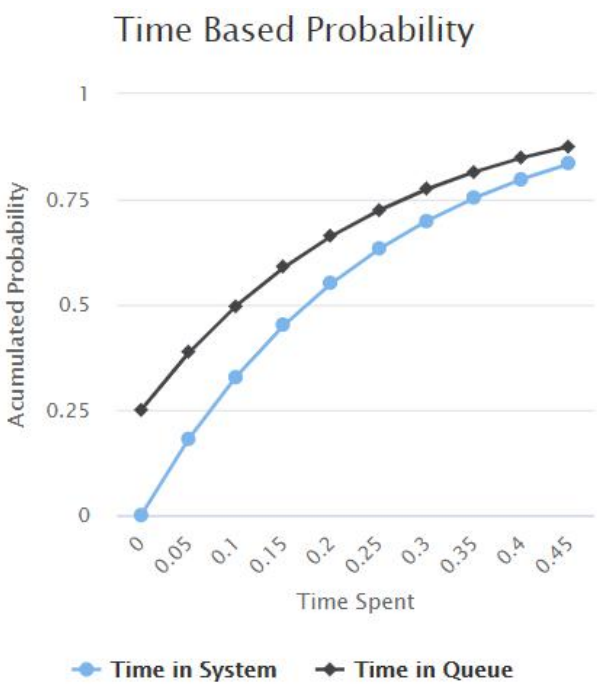
试着抽象描述

https://www.supositorio.com/rcalc/rcalclite.htm

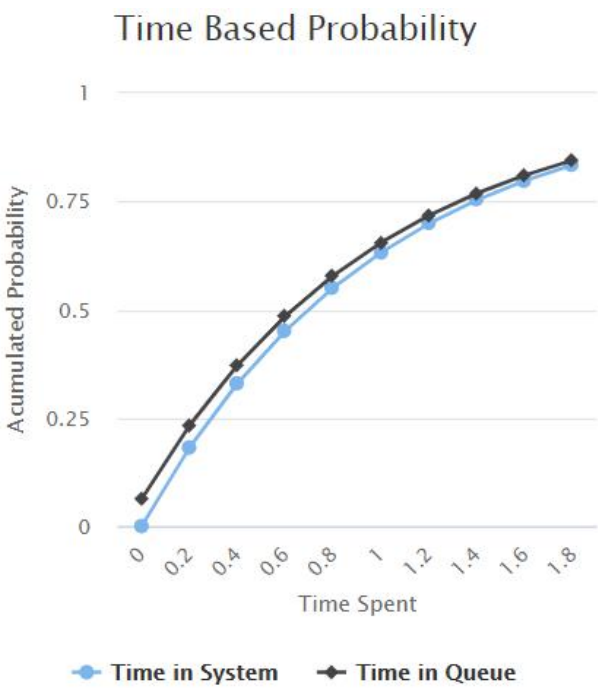
负载与尾延迟分布的关系



8/16/1



12/16/1



15/16/1

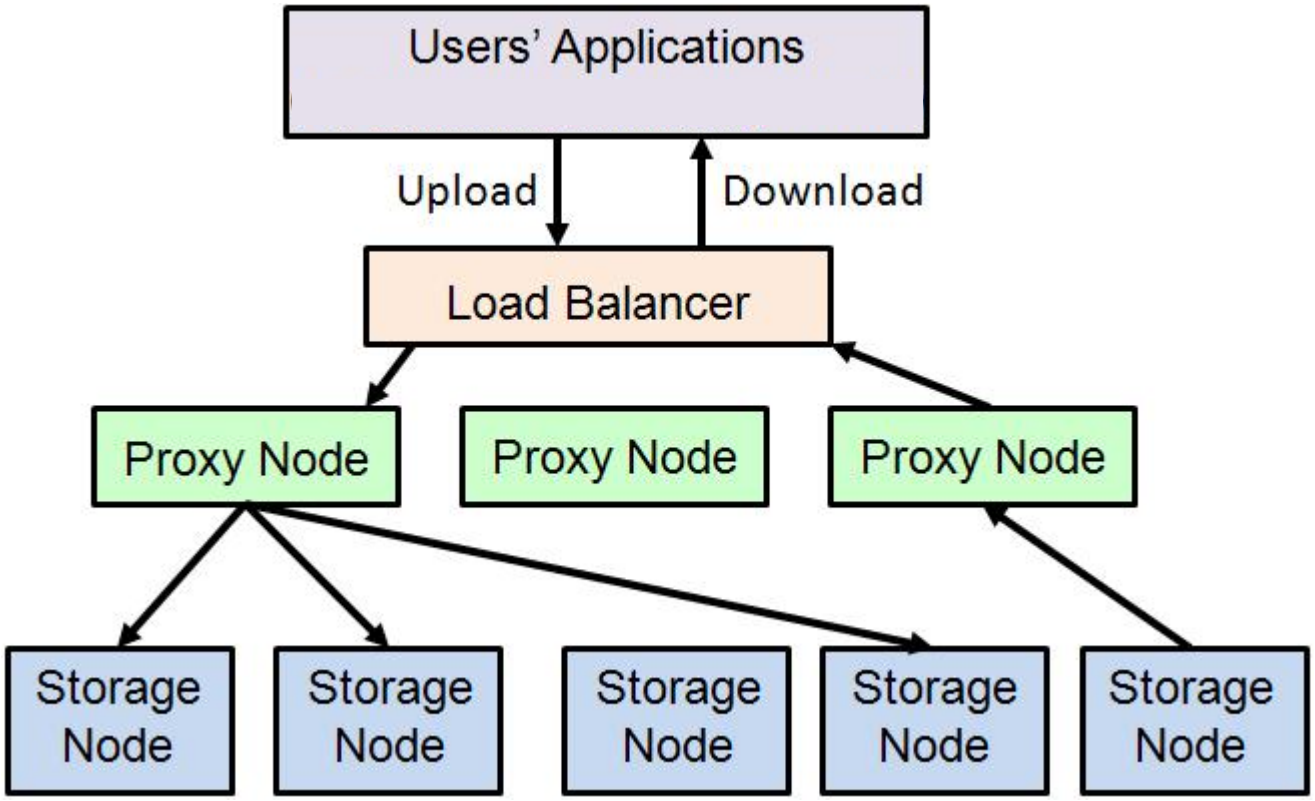
实际观测

```
s3bench \
        -accessKey=hust -accessSecret=hust2019 \
        -bucket=loadgen \
        -endpoint=http://192.168.3.85:9000 \
        -numClients=$1 \ {1, 2, 4, 8, 16, 32}
        -numSamples=1024 \
        -objectNamePrefix=loadgen \
        -objectSize=$(( 1024 * 256 ))
```

下面我们尝试分析现实系统

https://grafana.wikimedia.org/d/OPgmB1Eiz/swift?orgId=1

下面我们尝试分析现实系统



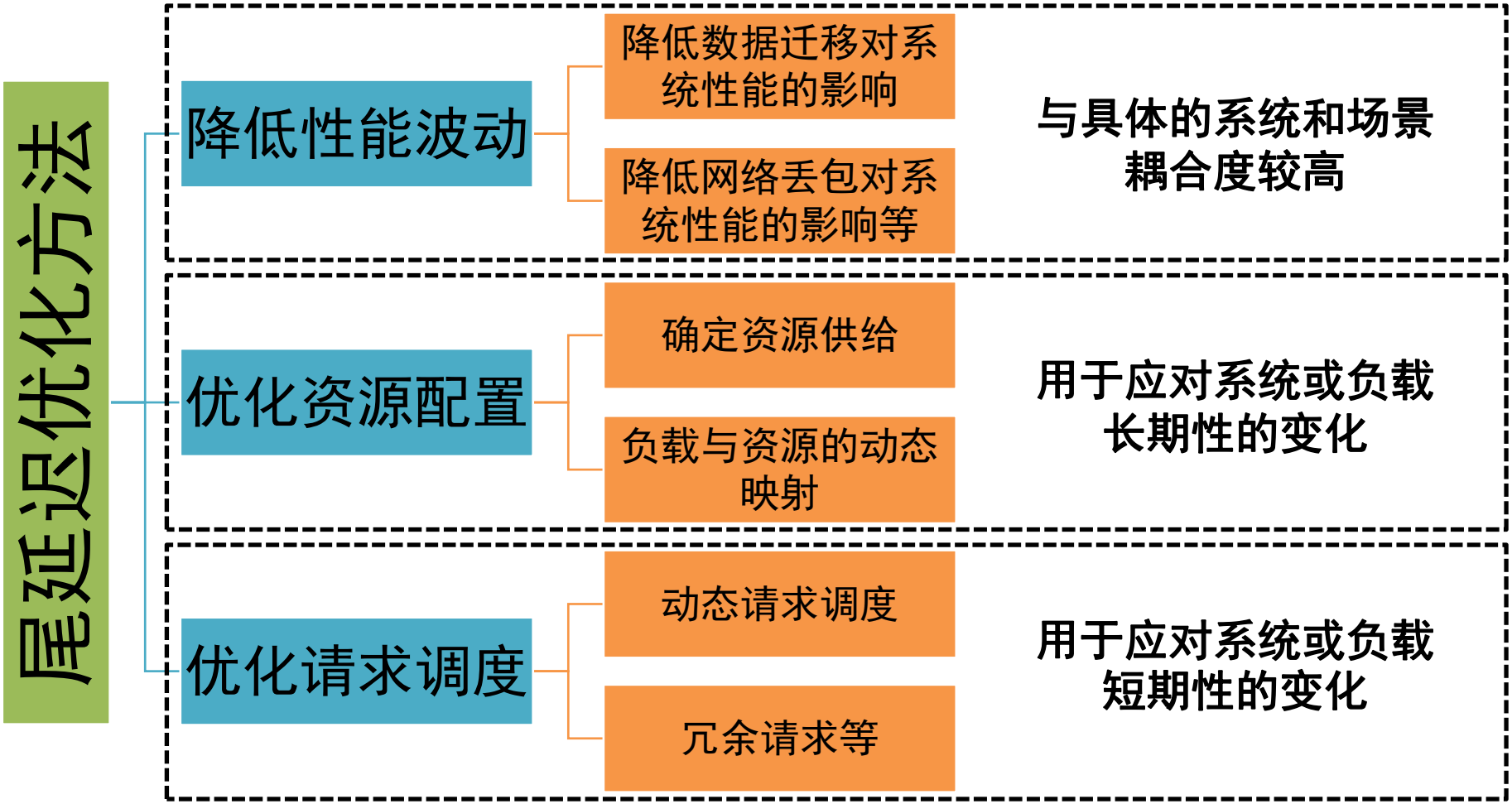
尾延迟优化的挑战

确定影响因素	适应动态变化	满足成本限制
需要对具体的系统、负载、场景进行分析，以确定会对尾延迟产生显著影响的因素	需要适应多种影响尾延迟的因素不可预测的动态变化	需要在针对会显著影响尾延迟的因素进行优化的同时，尽可能降低成本

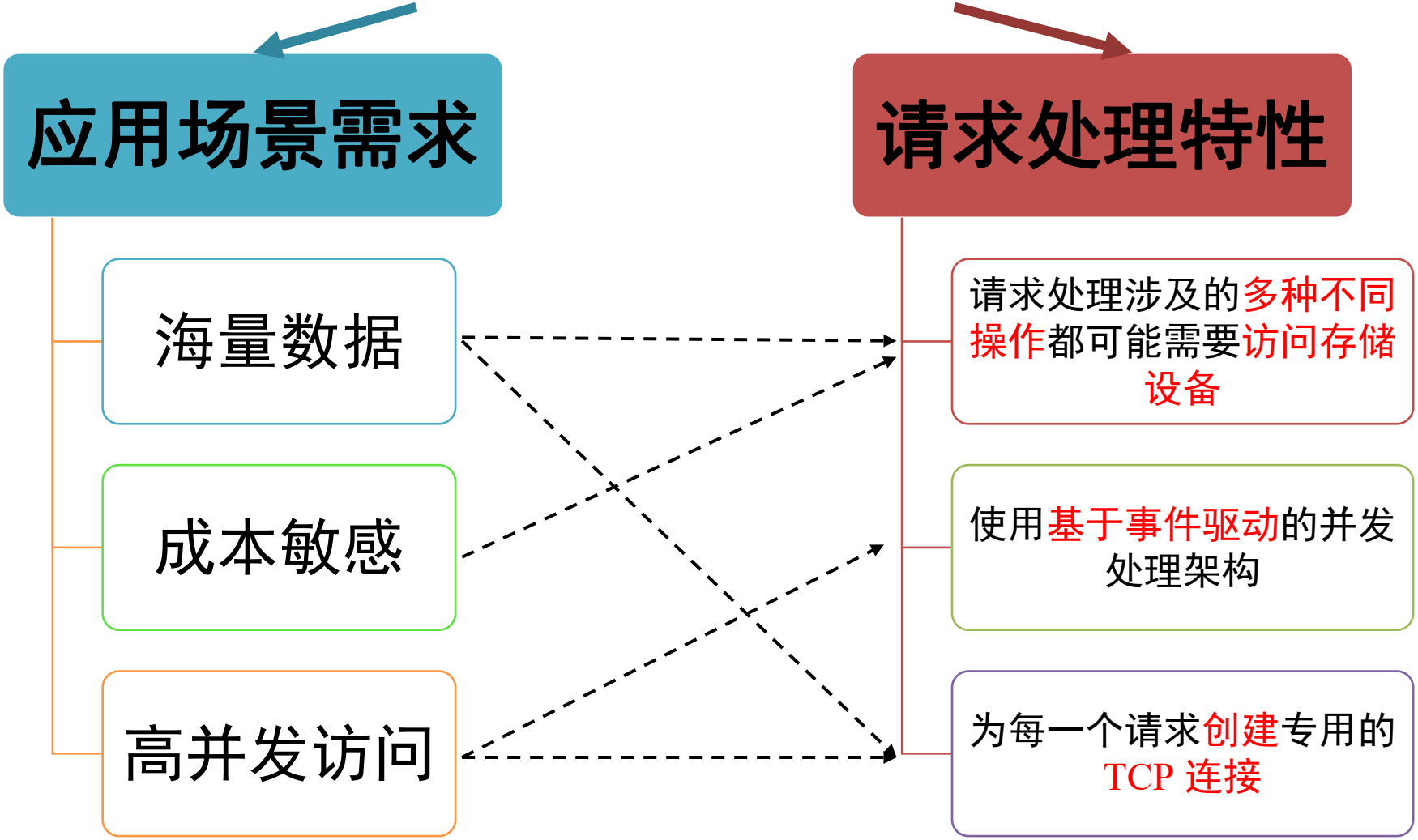
设备性能不稳定
设备失效
网络拥塞
网络丢包
网络多路径
系统的复杂机制
资源竞争
后台守护进程
调度策略的局限性
负载偏斜性
负载突发性
负载异构性
...

影响系统尾延迟的因素

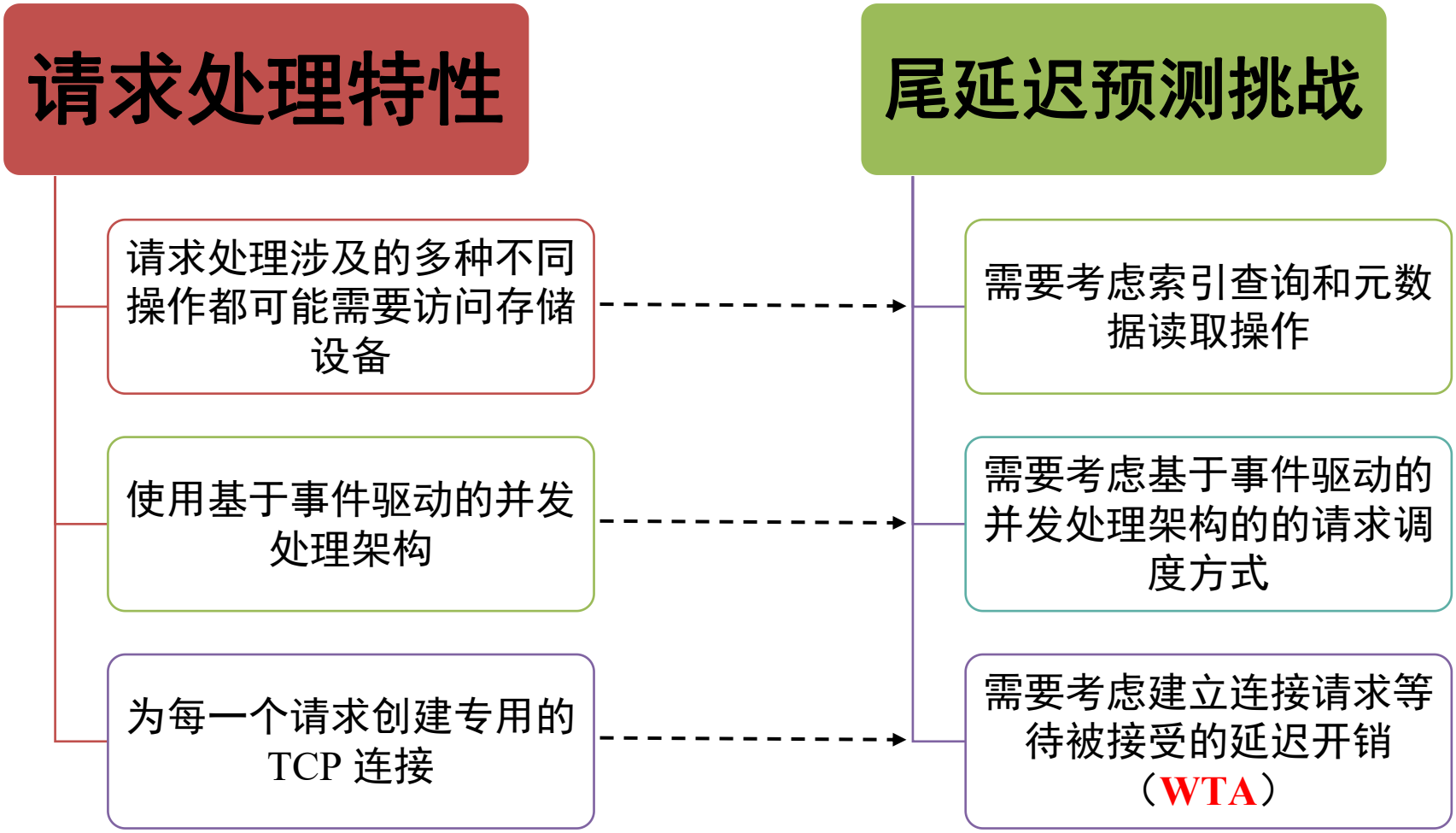
尾延迟优化的现状


尾延迟优化方法
降低性能波动
降低数据迁移对系统性能的影响
降低网络丢包对系统性能的影响等
与具体的系统和场景耦合度较高
优化资源配置
确定资源供给
负载与资源的动态映射
用于应对系统或负载长期性的变化
优化请求调度
动态请求调度
冗余请求等
用于应对系统或负载短期性的变化

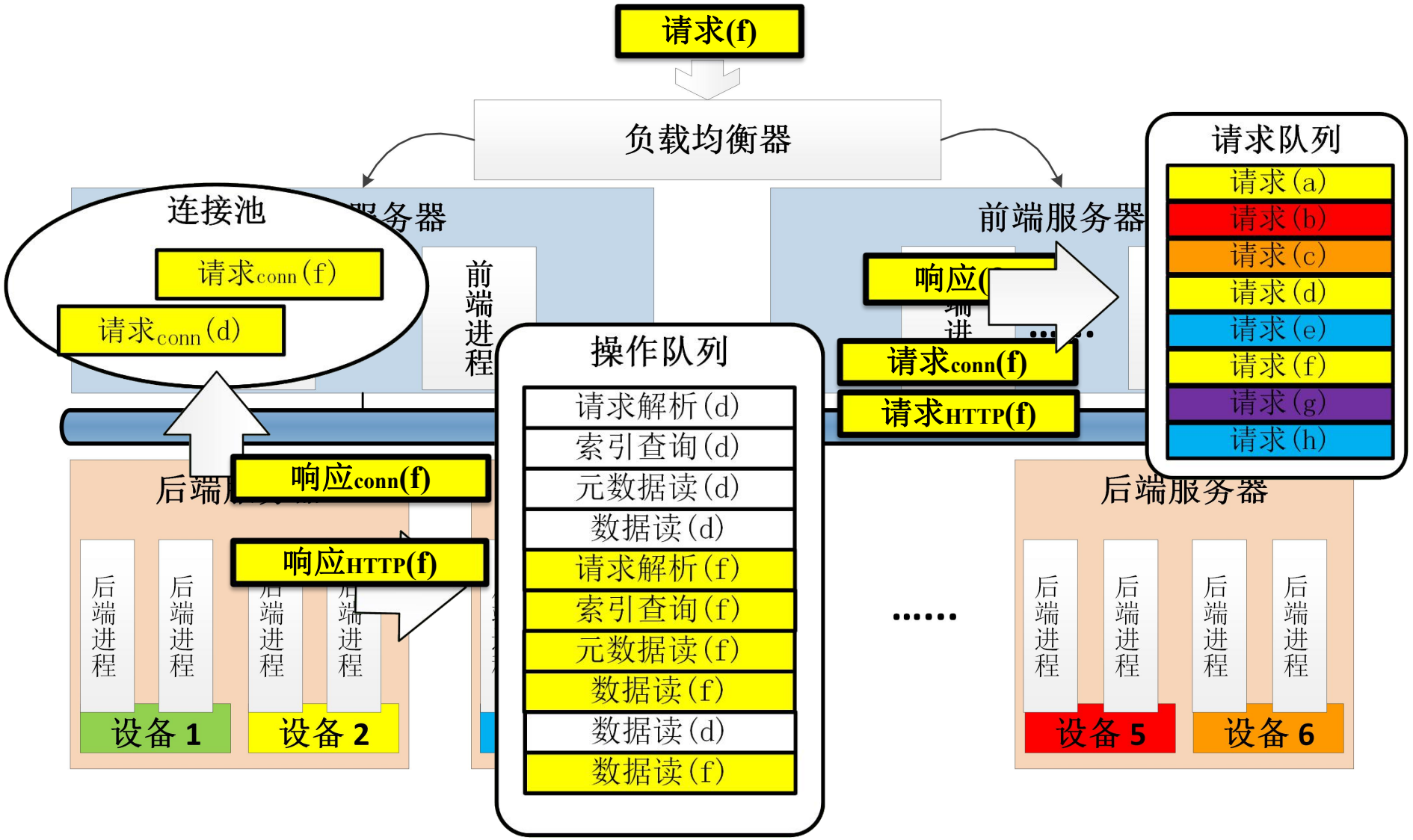
面向云应用的对象存储系统



尾延迟预测的挑战


请求处理特性
尾延迟预测挑战
请求处理涉及的多种不同操作都可能需要访问存储设备
使用基于事件驱动的并发处理架构
为每一个请求创建专用的 TCP 连接
需要考虑索引查询和元数据读取操作
需要考虑基于事件驱动的并发处理架构的的请求调度方式
需要考虑建立连接请求等待被接受的延迟开销（WTA）

请求处理过程


请求(f)
负载均衡器
请求队列
请求(a)
请求(b)
请求(c)
请求(d)
请求(e)
请求(f)
请求(g)
请求(h)
连接池
请求conn(f)
请求conn(d)
前端进程
前端服务器
响应(
请求conn(f)
请求HTTP(f)
操作队列
请求解析(d)
索引查询(d)
元数据读(d)
数据读(d)
请求解析(f)
索引查询(f)
元数据读(f)
数据读(f)
数据读(d)
数据读(f)
响应conn(f)
响应HTTP(f)
后端进程
设备 1
设备 2
设备 5
设备 6
后端服务器

COSModel性能模型

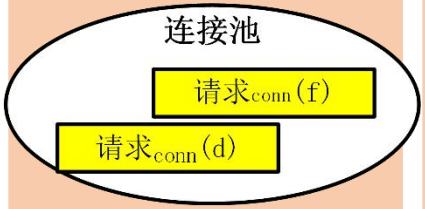
系统整体的延迟分布

一个存储设备在前端层的延迟分布

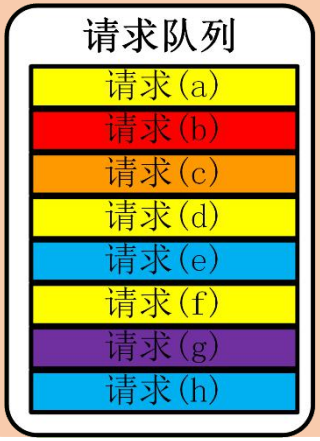
……

一个存储设备在后端层的延迟分布

WTA 的分布



请求在前端层中的排队延迟分布



COSModel - 后端层

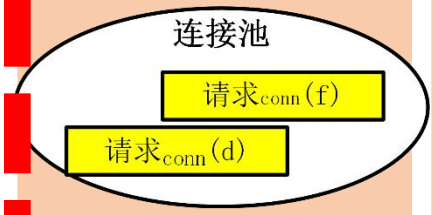
系统整体的延迟分布

一个存储设备在前端层的延迟分布

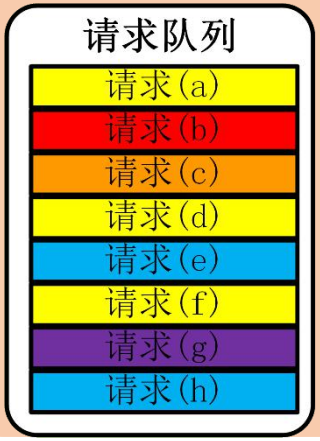
……

一个存储设备在后端层的延迟分布

WTA 的分布

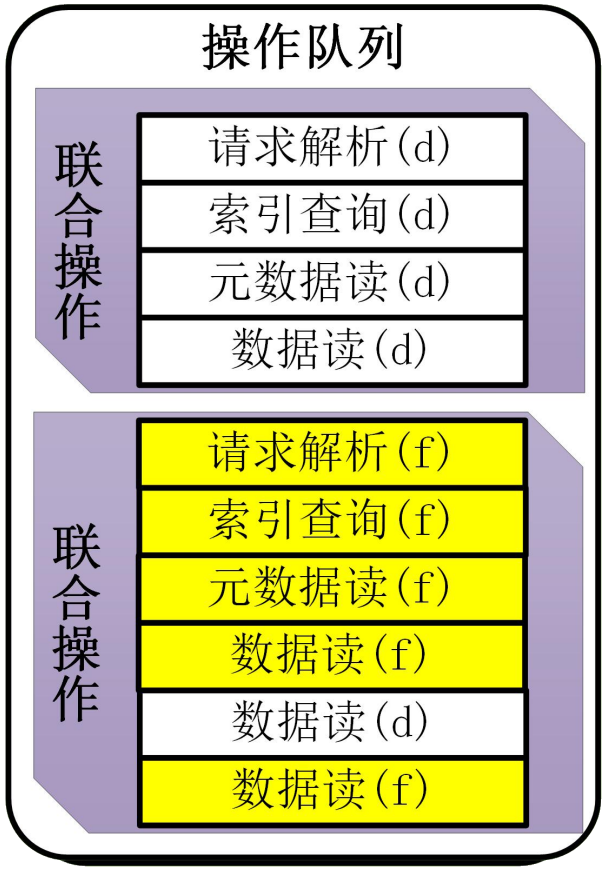


请求在前端层中的排队延迟分布



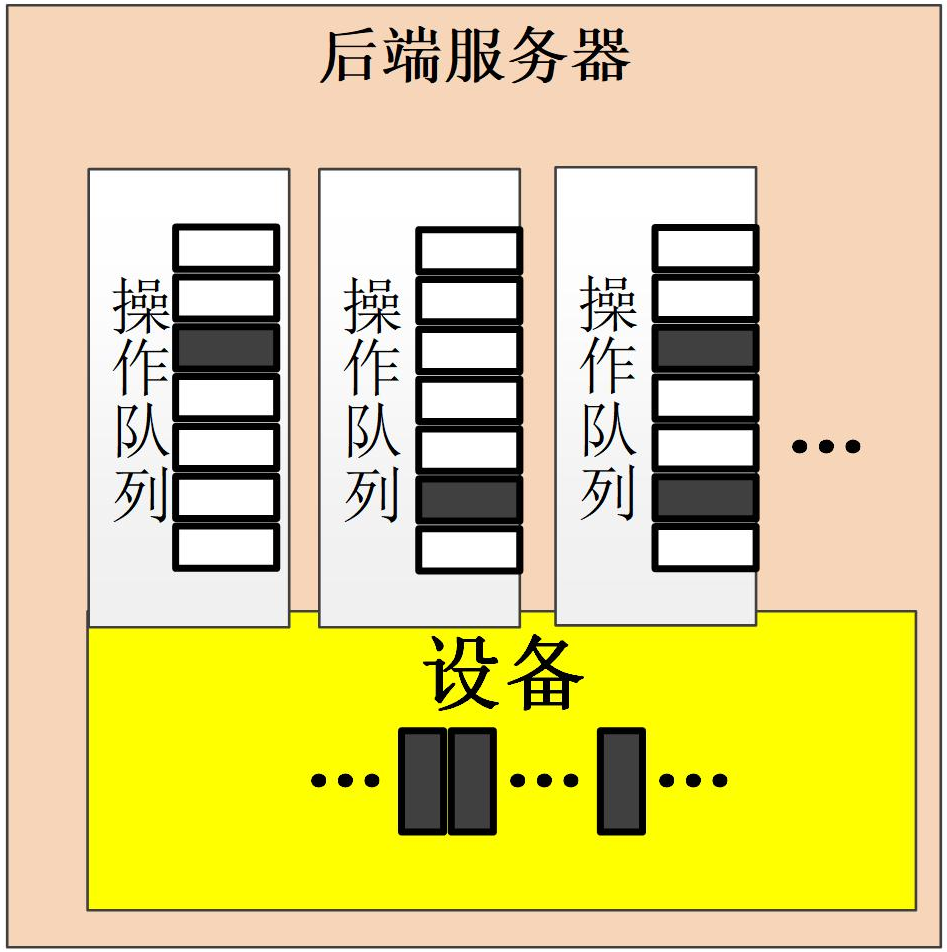
COSModel - 后端层

- 将多个操作打包为联合操作
  - 一个**请求解析**操作及其后的**非请求解析**操作
- 使用 M/G/1 队列模型为联合操作队列建模

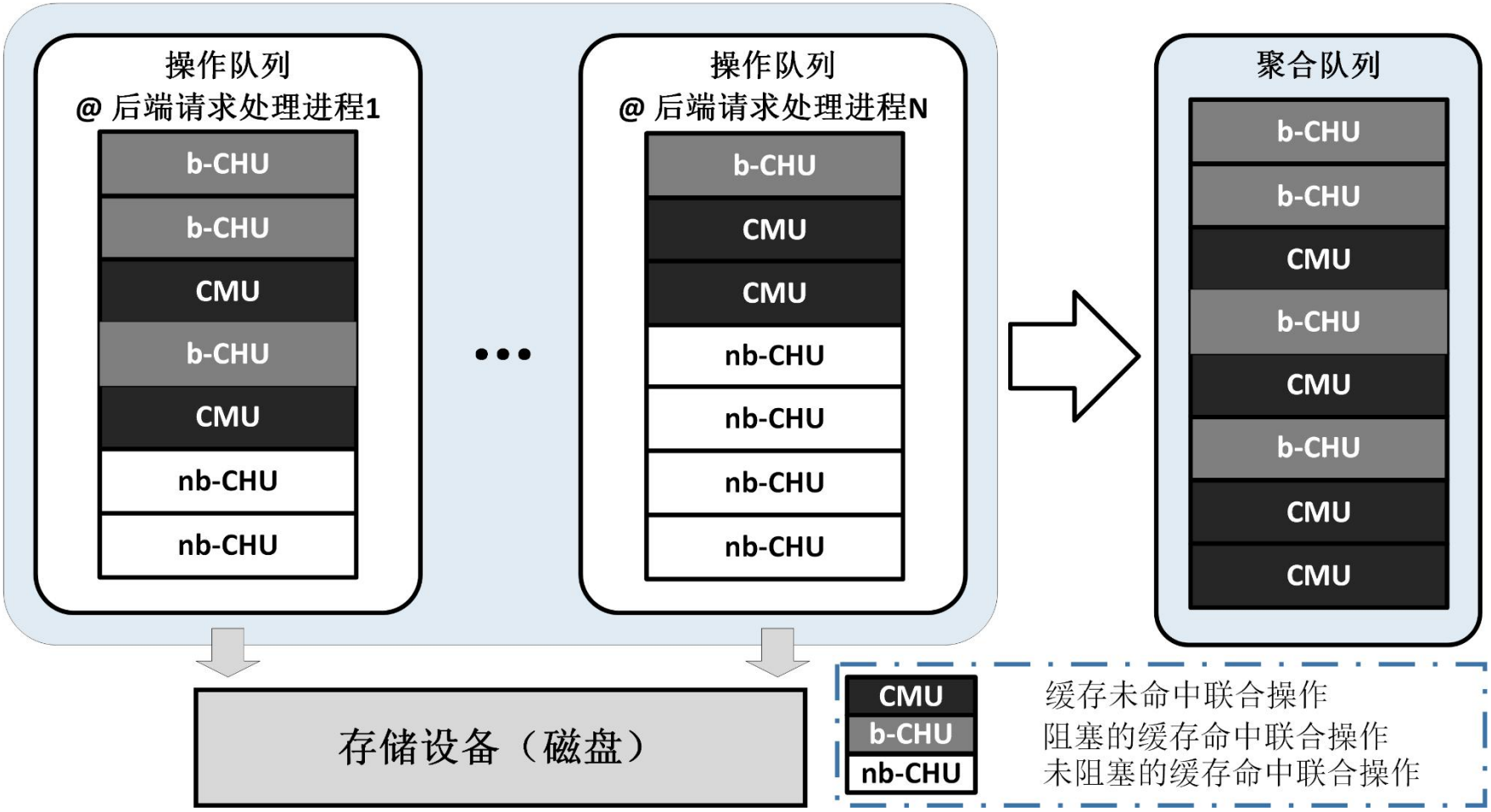


COSModel - 后端层

- 一个存储设备对应于多个独立的后端进程
  - 多个相互影响的操作队列
- 使用一个聚合队列近似多个操作队列



COSModel - 后端层

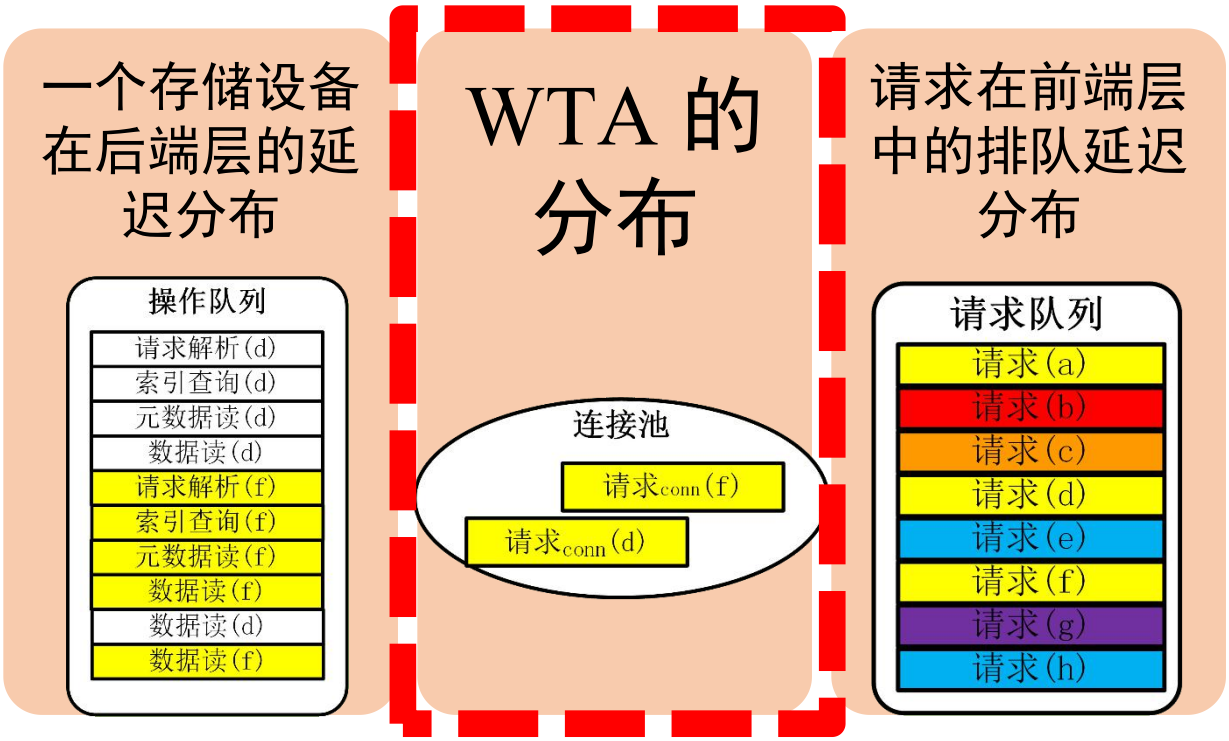


COSModel - WTA

系统整体的延迟分布

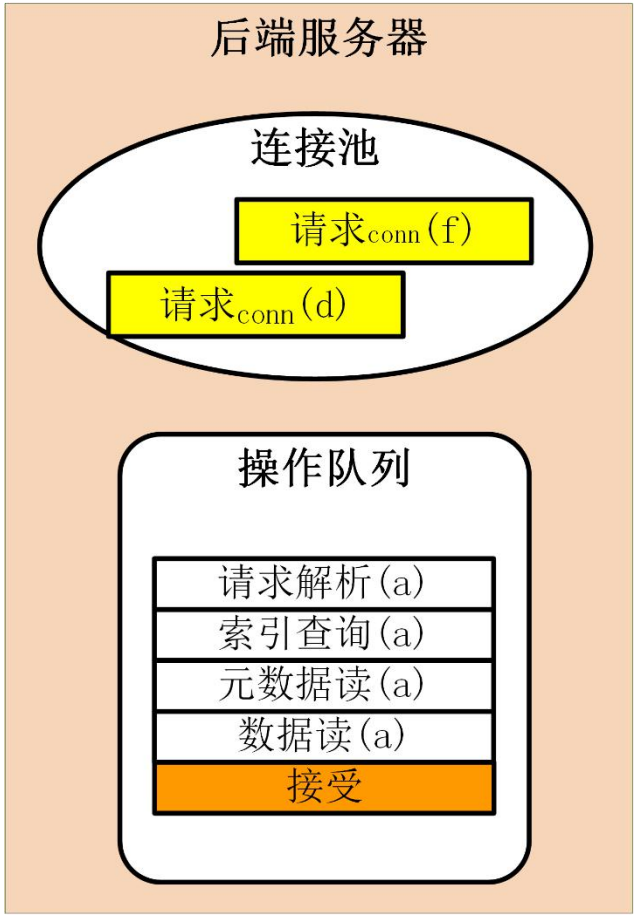
一个存储设备在前端层的延迟分布

……



COSModel - WTA

- 一个建立连接请求的 WTA 与接受操作在操作队列中的等待时间相关
- 使用操作队列的等待时间分布近似WTA的分布



COSModel – 前端层

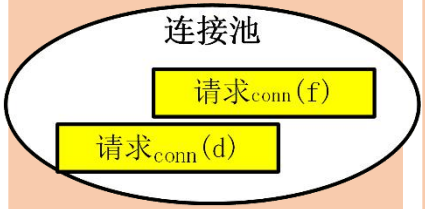
系统整体的延迟分布

一个存储设备在前端层的延迟分布

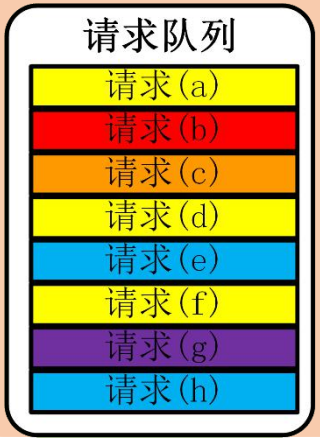
……

一个存储设备在后端层的延迟分布

WTA 的分布

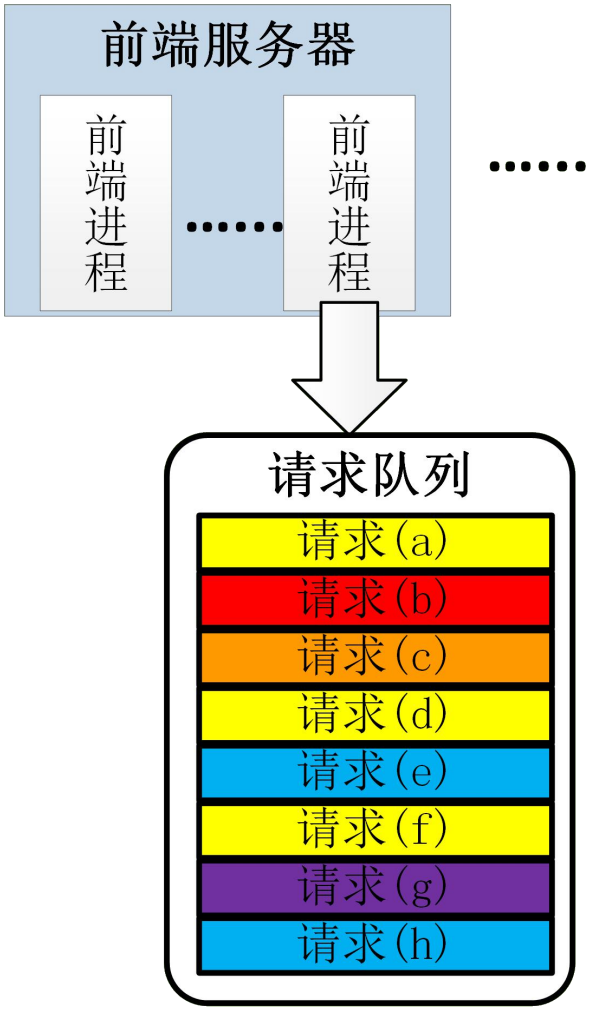


请求在前端层中的排队延迟分布



COSModel – 前端层

- 对等的前端进程
  - 前端层整体的排队时延与一个前端进程的排队时延相同
- 使用 M/G/1 队列模型对前端进程的请求队列建模



COSModel – 前端层

- 一个存储设备在前端层上的响应时延包含三个部分:
  - 这个存储设备在后端层的响应时延
  - WTA
  - 请求在前端层的排队时延
- 使用卷积将各个部分的时延结合起来

COSModel – 系统整体

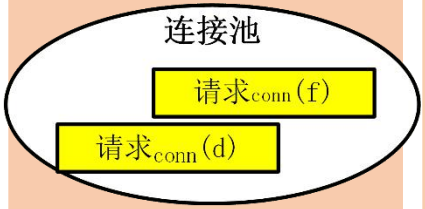
系统整体的延迟分布

一个存储设备在前端层的延迟分布

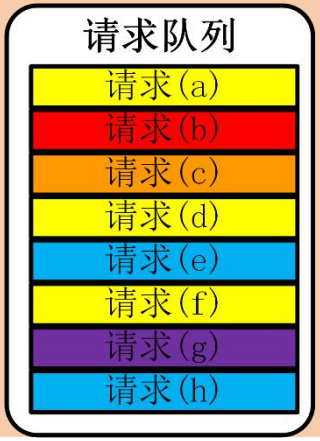
……

一个存储设备在后端层的延迟分布

WTA 的分布



请求在前端层中的排队延迟分布



COSModel – 系统整体

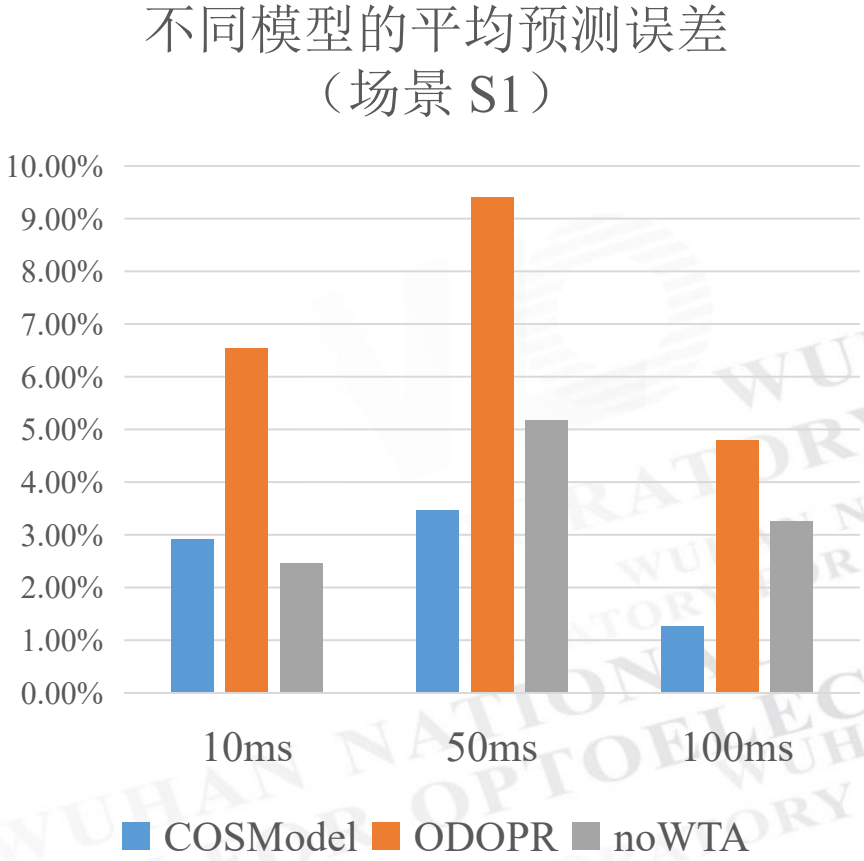
- 系统整体的时延分布是一个混合分布
  - 混合部分
    - 各个存储设备在前端层的响应时延分布
  - 混合权重
    - 存储设备的负载所占总体负载的比例

准确性评估 - 实验设置

- **实验平台**
  - 一个 OpenStack Swift 集群 （7个节点）
- **数据集**
  - 维基百科中多媒体文件的访问（WikiBench）
- **响应时延要求（SLA）**
  - 10ms, 50ms, 100ms
- **基准模型**
  - ODOPR
    - 假设请求处理过程中索引查询，元数据读操作都缓存命中，无需访问存储设备
  - noWTA
    - 假设请求不存在等待被接受的时间（WTA）

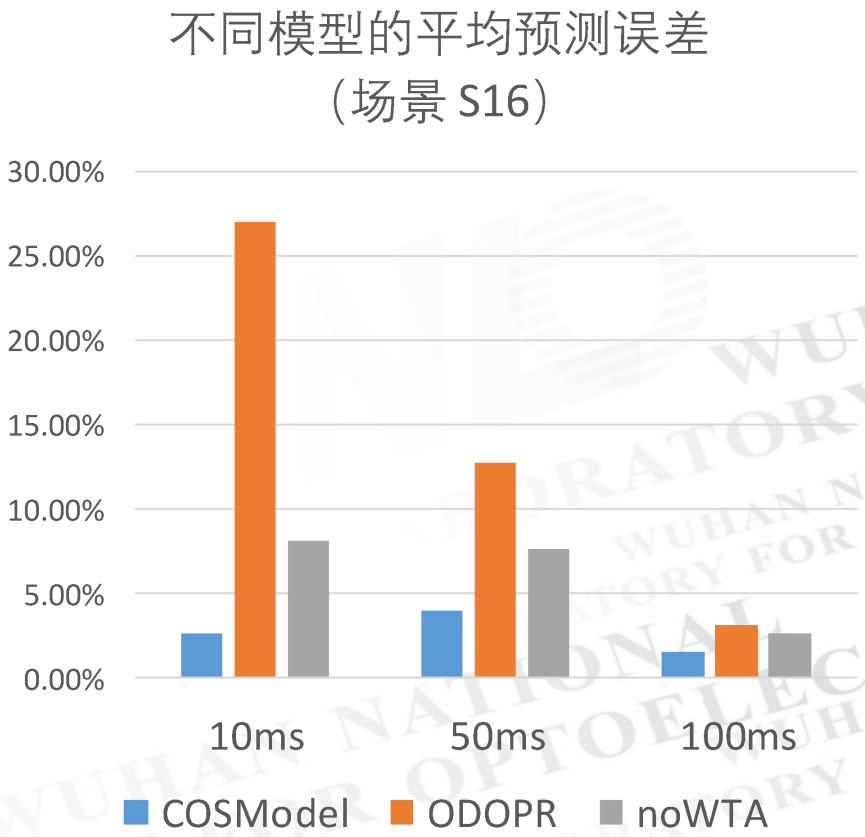
准确性评估 – 单后端进程

- **场景** S1
  - 1个存储设备对应1个后端进程
- **负载**
  - 从 10 个请求每秒到350个请求每秒，按 5 递增



准确性评估 – 多后端进程

- **场景** S16
  - 1个存储设备对应16个后端进程
- **负载**
  - 从 10 个请求每秒到600个请求每秒，按 5 递增



本讲小结

- 建立了一个基于分析的性能模型以预测云对象存储系统的尾响应延迟
- 主要贡献
  - 抽象出联合操作
  - 请求等待被接受延迟开销的量化分析
- 研究成果
  - Predicting Response Latency Percentiles for Cloud Object Storage Systems (ICPP 2017)
  - Understanding the latency distribution of cloud object storage systems (JPDC)