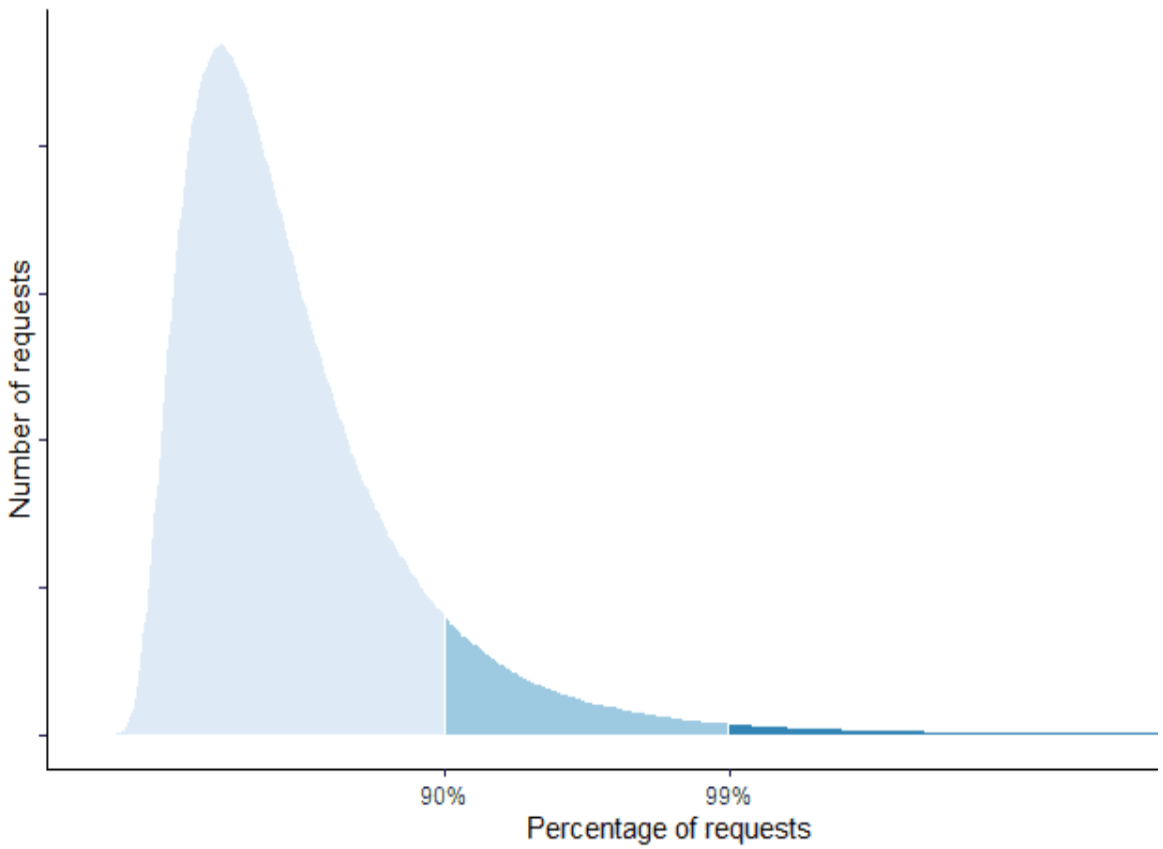
DATACENTER
REPORT

When Idling is Ideal: Optimizing Tail-Latency for Heavy-Tailed Datacenter Workloads with Perséphone

胡悦晨

2021/12/31

1.背景介绍



低利用率运行

浪费了CPU运行的时间

共享队列和工作窃取

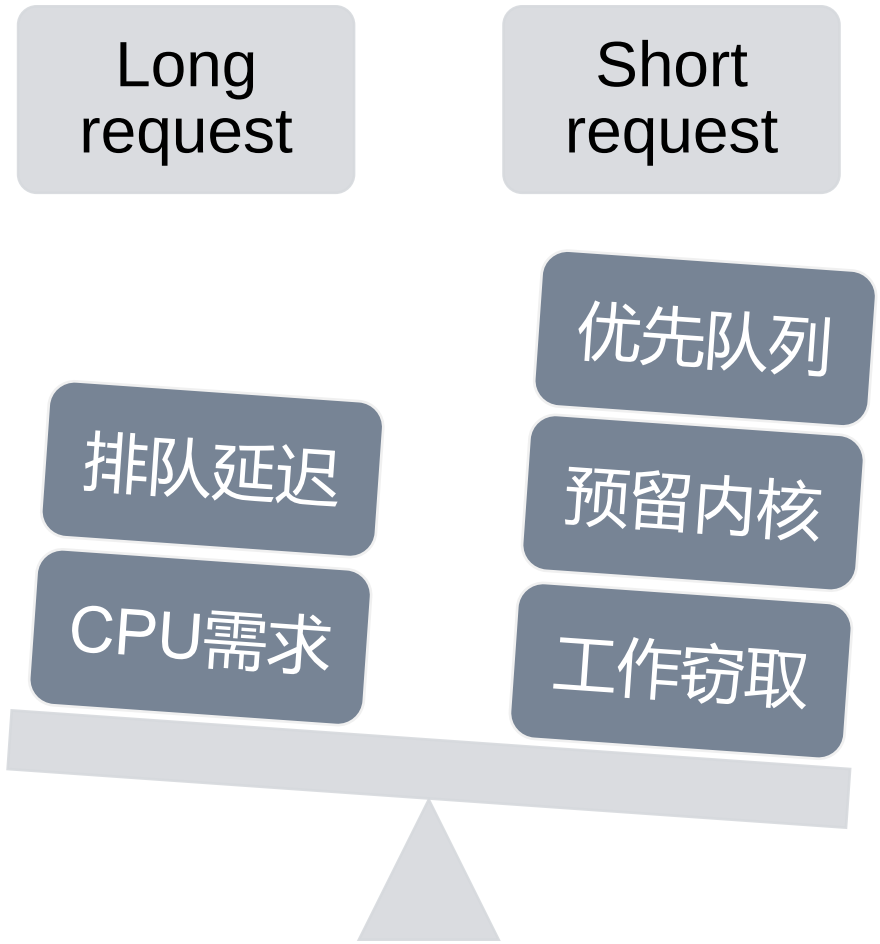
只适用于均匀和轻尾的工作负载

对短请求进行优先级排序

很难在微秒级实现

Perséphone整合了一个动态的、应用程序感知的预留核心(DARC)，它只对短请求利用工作保存。最适合那些看重微秒响应的应用程序

2.DARC模型



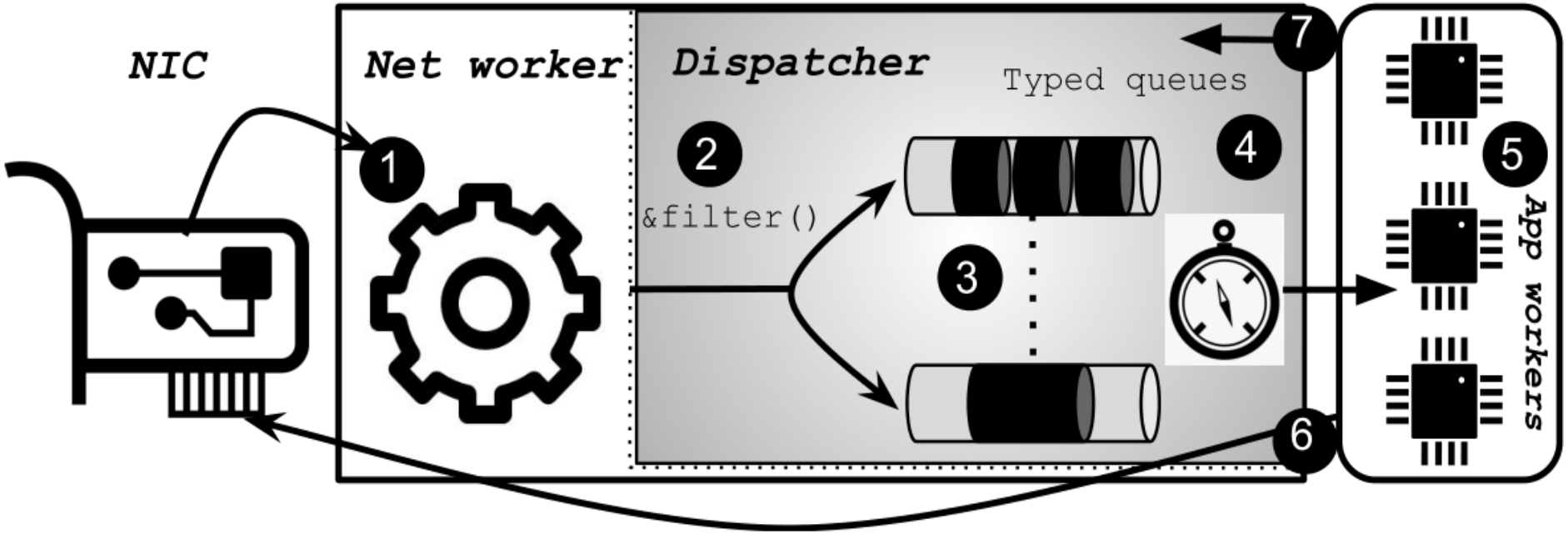
$$0 \leq \frac{S_i * R_i}{\sum_j^N S_j * R_j}, \leq 1$$

Algorithm 1 Request dispatching algorithm

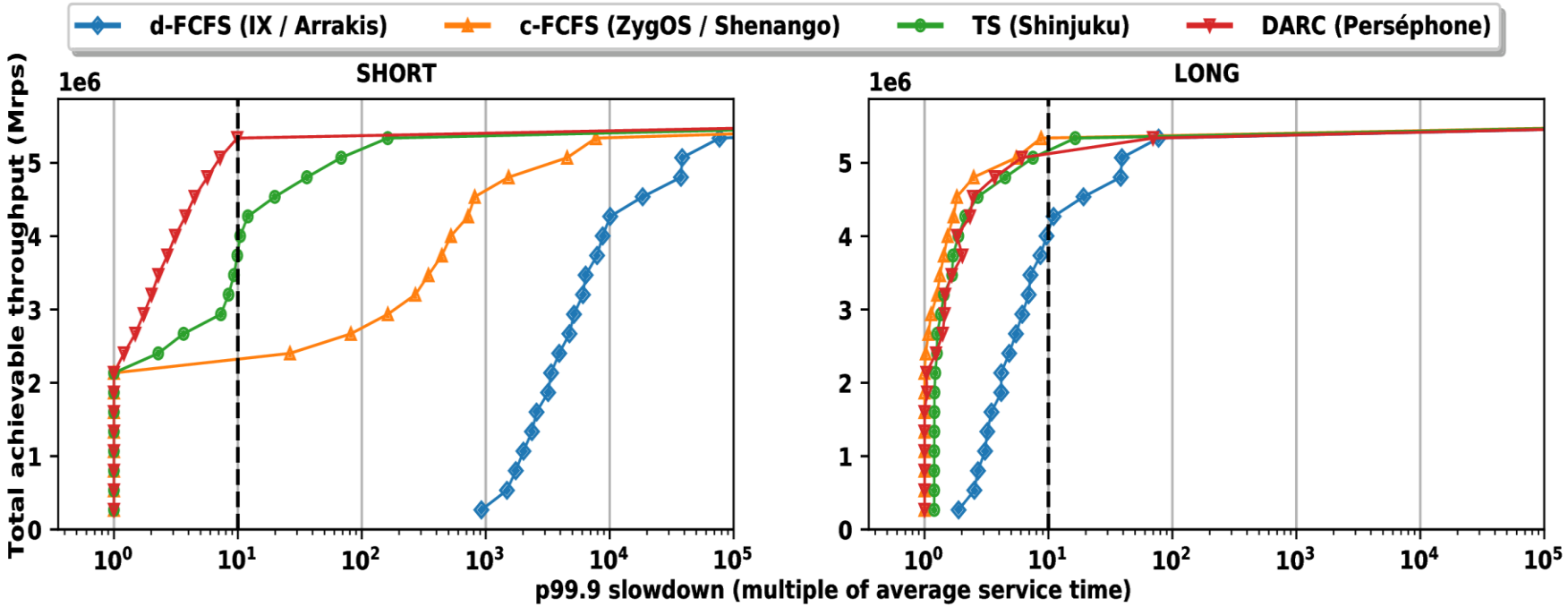
```
procedure DISPATCH(Types)
    w ← None
    for τ ∈ Types.sort() do
        if τ.queue == ∅ then
            continue
        else
            workers ← τ.reserved ∪ τ.stealable
            for worker ∈ workers do
                if worker.is_free() then
                    w ← worker
                    break
            if w ≠ None then
                r ← τ.queue.pop()
                schedule(r, w)
```

2. Perséphone架构

①网络工作端　　②调度器　　③应用程序端



3. 性能评估



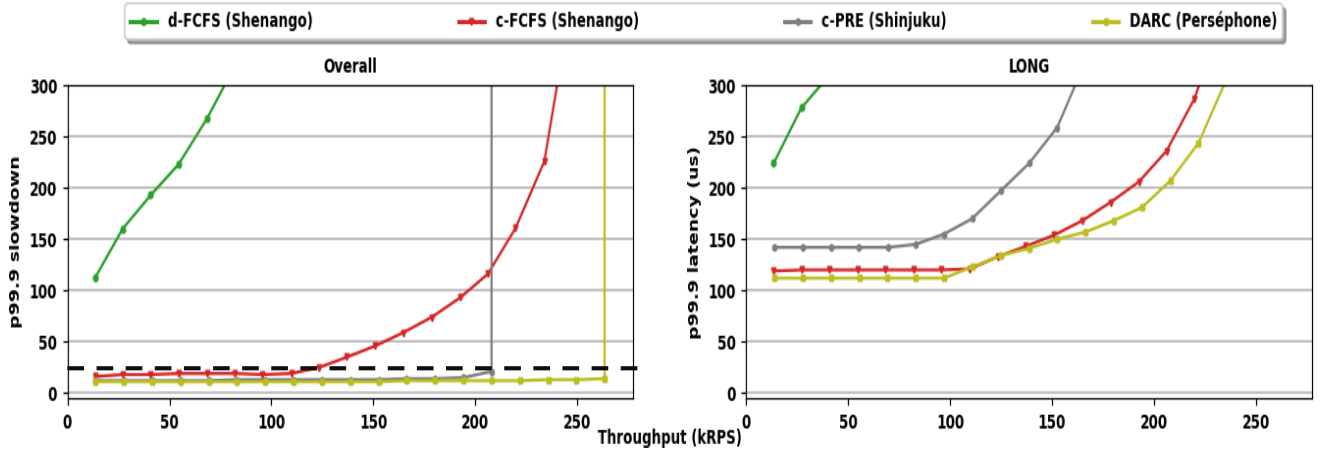
使用16核系统，当目标SLO为每种请求类型平均服务时间的10倍时，c-FCFS和TS分别只能处理2.1百万和3.7百万请求/秒(Mrps)。DARC可为同一目标维持5.1 Mrps。

3. 性能评估——负载离散度不同

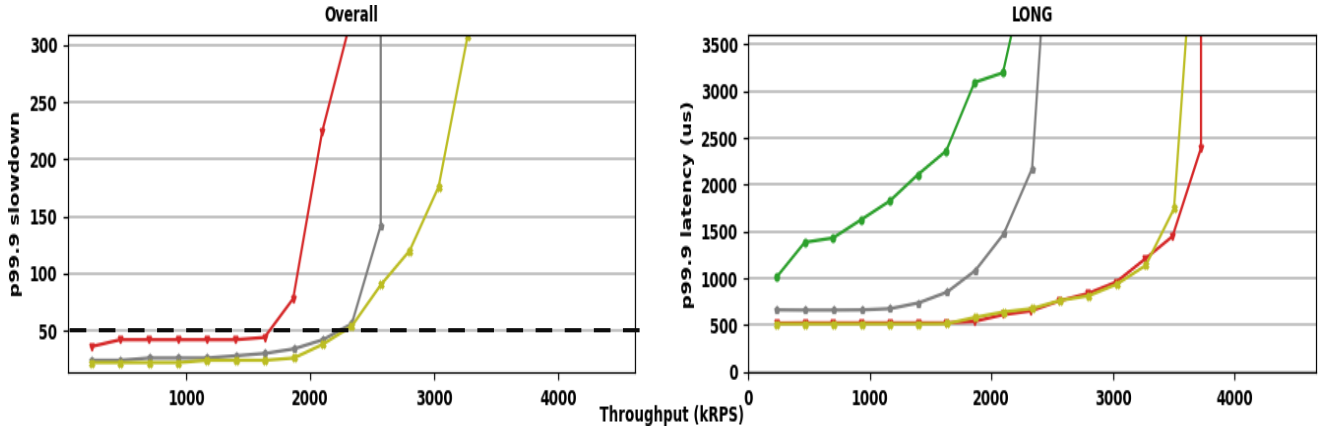
Workload	*Short* Runtime (μs)	Ratio	*Long* Runtime (μs)	Ratio
High Bimodal	1	50%	100	50%
Extreme Bimodal	0.5	99.5%	500	0.5%

离散度为100倍的工作负载，对于20倍的减速目标，DARC可以分别比Shenango和新宿多维持**2.35倍**和**1.3倍**的流量

离散度为1000倍的工作负载， Perséphone可以比Shenango多维持**1.4倍**的吞吐量，并在短请求方面比新宿提高**1.4倍**的速度

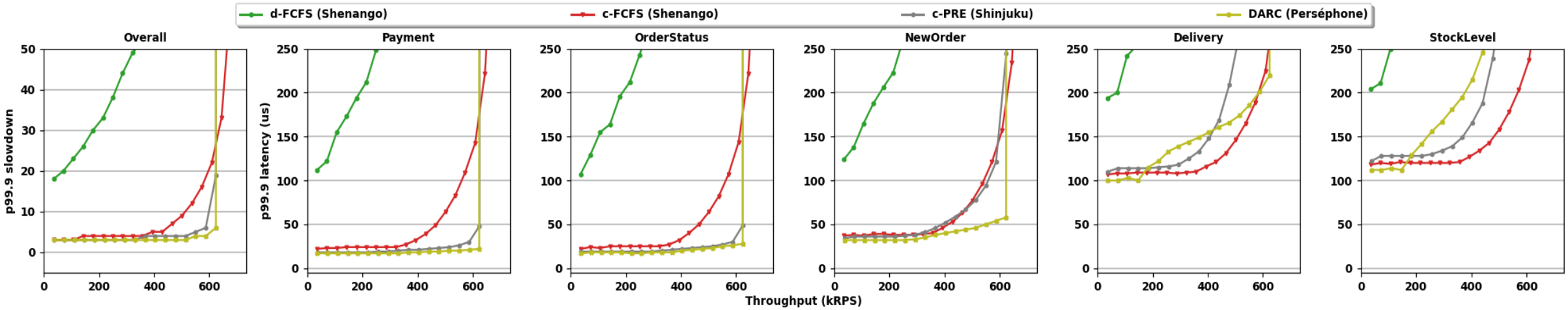


High Bimodal



Extreme Bimodal

3. 性能评估



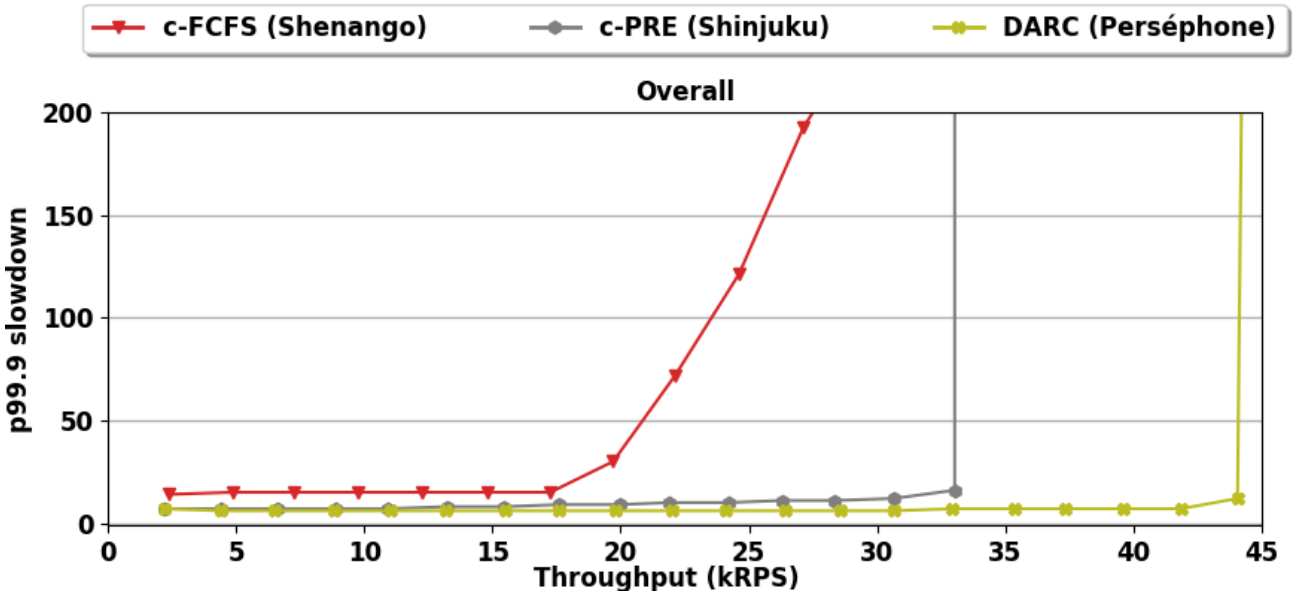
Transaction name	Runtime (μs)	Ratio	Dispersion
Payment	5.7	44%	1x
OrderStatus	6	4%	1.05x
NewOrder	20	44%	3.3x
Delivery	88	4%	15.4x
StockLevel	100	4%	17.5x

在线商店模型测试（以TPC-C基准）

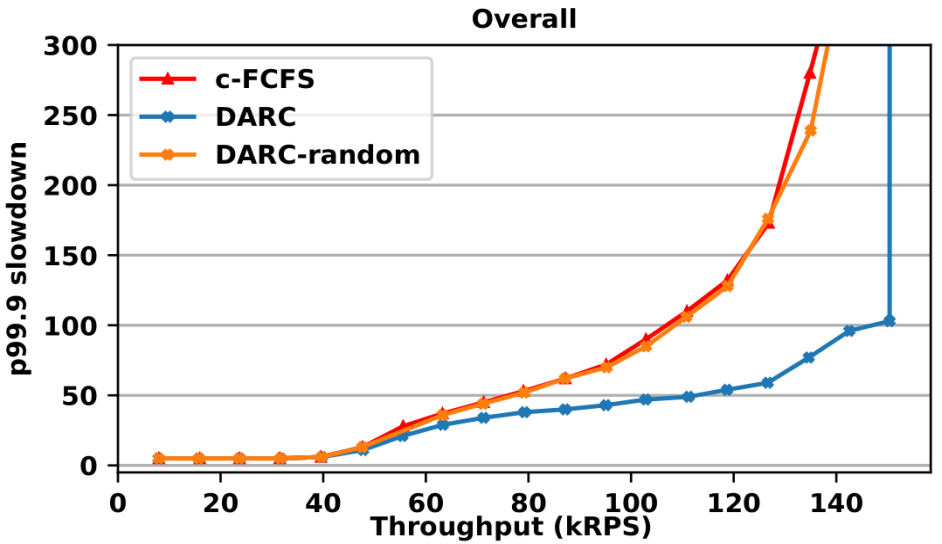
与Shenango的c-FCFS相比，DARC为**支付、订单状态和新委托单量事务分别提供高达9.2倍、7倍和3.6倍的吞吐量。**

代价是来自较长请求的投递和存储级事务遭受了更高的尾延迟。

3. 性能评估



RocksDB (Facebook使用的数据库引擎)对于20倍的减速目标，DARC可以比Shenango和新宿分别维持2.3倍和1.3倍的高吞吐量。



评估当用户不能提供正确的请求分类器时DARC的行为，DARC-random的行为收敛于c-FCFS

4. 总结与展望

1.网络模型瓶颈

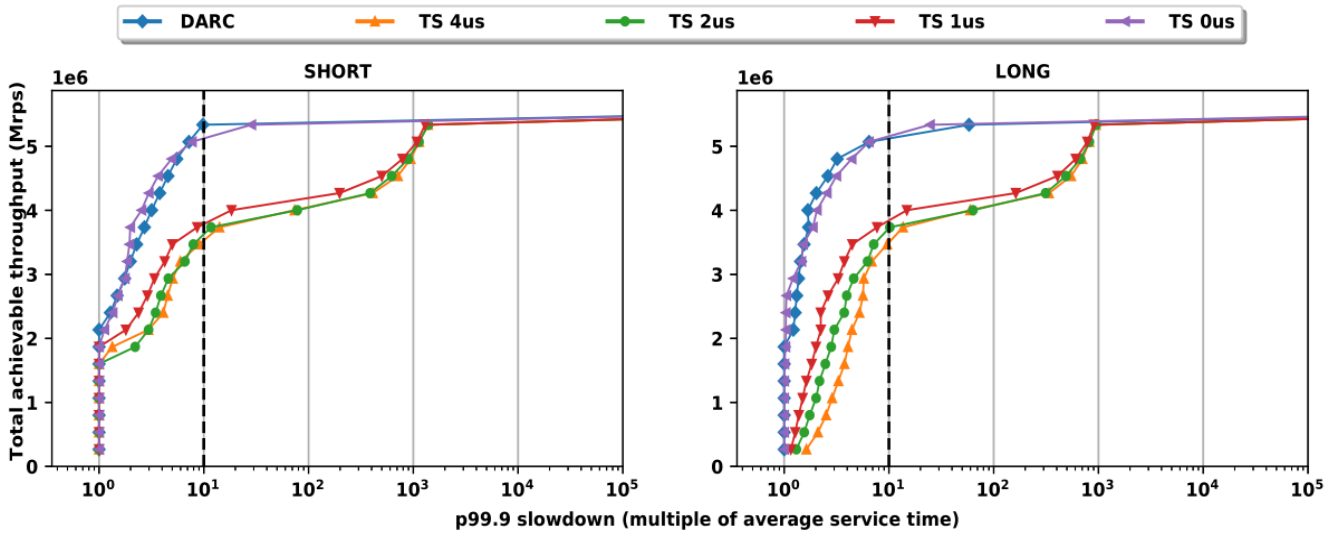
网络工作端是一个第2层转发器，并对以太网和IP报头执行简单的检查；应用程序端处理第4层及以上的转发器，并直接执行TX。该设计旨在最大化调度程序的性能，但也Perséphone的主要瓶颈，并使其与现有系统竞争。

2.在微秒尺度上，实现和协调抢占式系统仍具有挑战性

在微秒级下，对于短请求，放缓目标为10，即使是1微秒的开销也会导致30%可持续负载减少

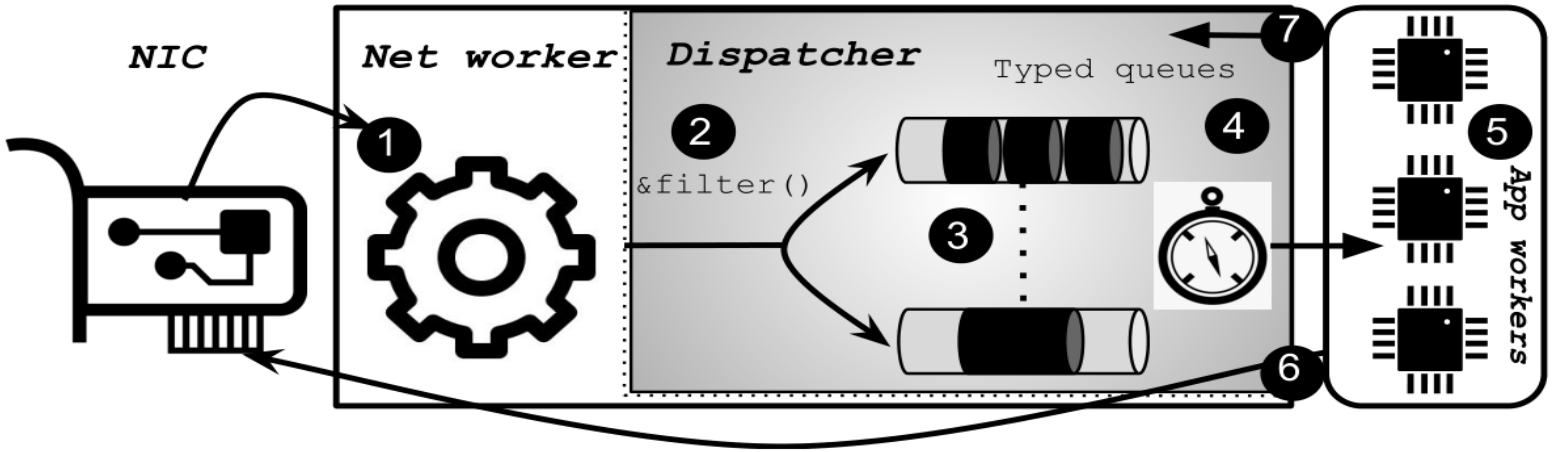
3.数据中心系统中的DARC

DARC可以与分配器合作获取和释放核心，适应负载变化，并在此类事件中更新保留



4. 总结与展望

本文提出了一个新的内核旁路操作系统调度器Perséphone，它实现了**应用程序感知，非工作保存的DARC策略**。在重尾工作负载中，DARC**将核心用于短请求，**以保证它们不会被长请求阻塞，从而为较短的请求**保持了良好的尾延迟**。
我们的Perséphone原型为更短的请求保持了良好的尾延迟，并且与最先进的内核旁路调度程序相比，可以**用相同数量的内核处理更高的负载**，总体上**更好地利用了数据中心资源**。



感谢聆听

胡悦晨

2021/12/31