第三部分

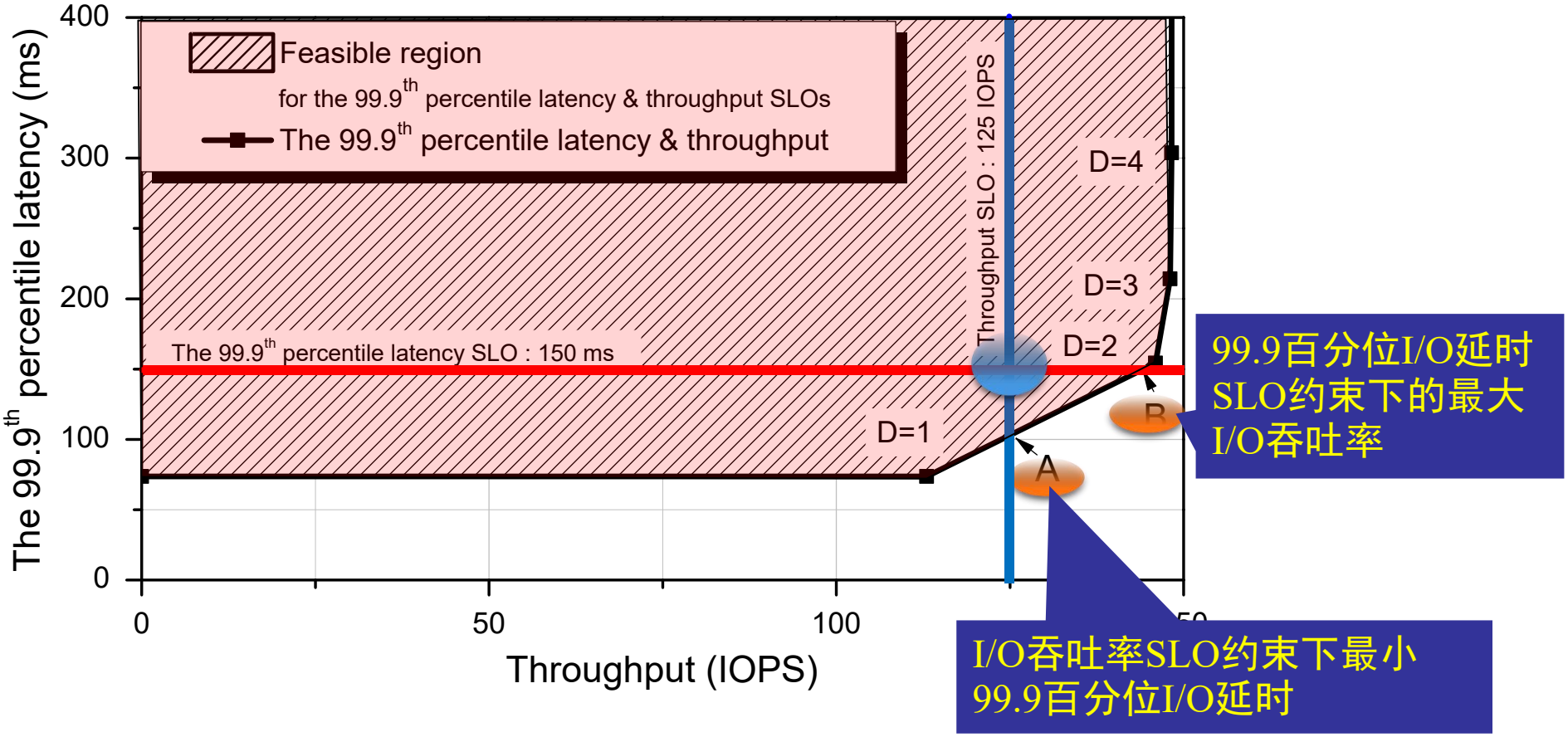
两维SLO约束的I/O优化

PSLO: enforcing the Xth percentile latency and throughput SLOs for consolidated VM storage[C]//Proceedings of the Eleventh European Conference on Computer Systems. 2016

研究现状

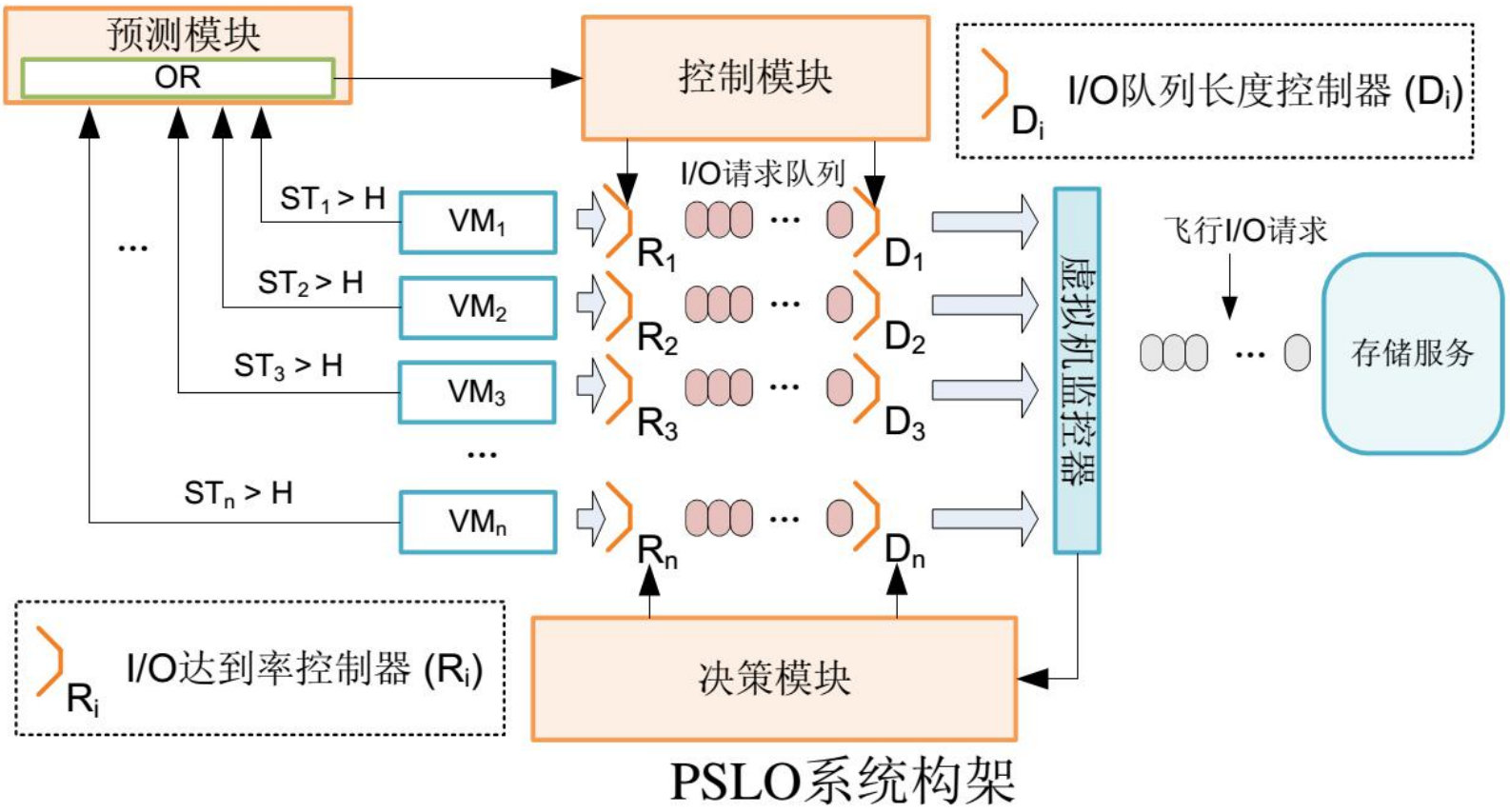
- 性能保障和优化并不孤立，而是相辅相成的概念，即保障是优化的重要前提，而优化是基于保障约束的资源利用最大化。
- 现有的存储性能保障和优化的相关研究主要面向I/O 吞吐率、I/O 带宽和I/O 延时这三类性能指标，未能充分考虑不同性能指标SLO保障的隔离，以及性能优化行为与SLO保障行为的隔离。
- 所以，现有方案难以有效实现多性能指标SLO精确保障，想进一步挖掘现有资源优化性能自然就难上加难。

研究动机



通过控制I/O并行度优化既定百分位I/O延时或者I/O吞吐率。

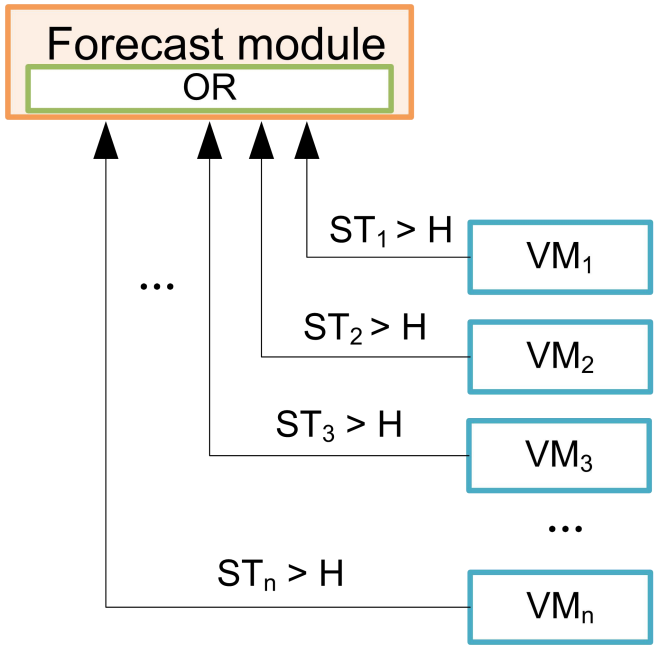
总体设计



PSLO系统构架

PSLO由三个关键功能模块、预测模块、控制模块和决策模块组成。该系统可以根据用户的具体需要设定每个虚拟机必须满足的百分位尾延时SLO和吞吐率SLO。然后PSLO可以根据不同的SLO策略进行这两类SLO的保障和性能优化。

预测模块



n个并行虚拟机发出SLO违例预测的概率：

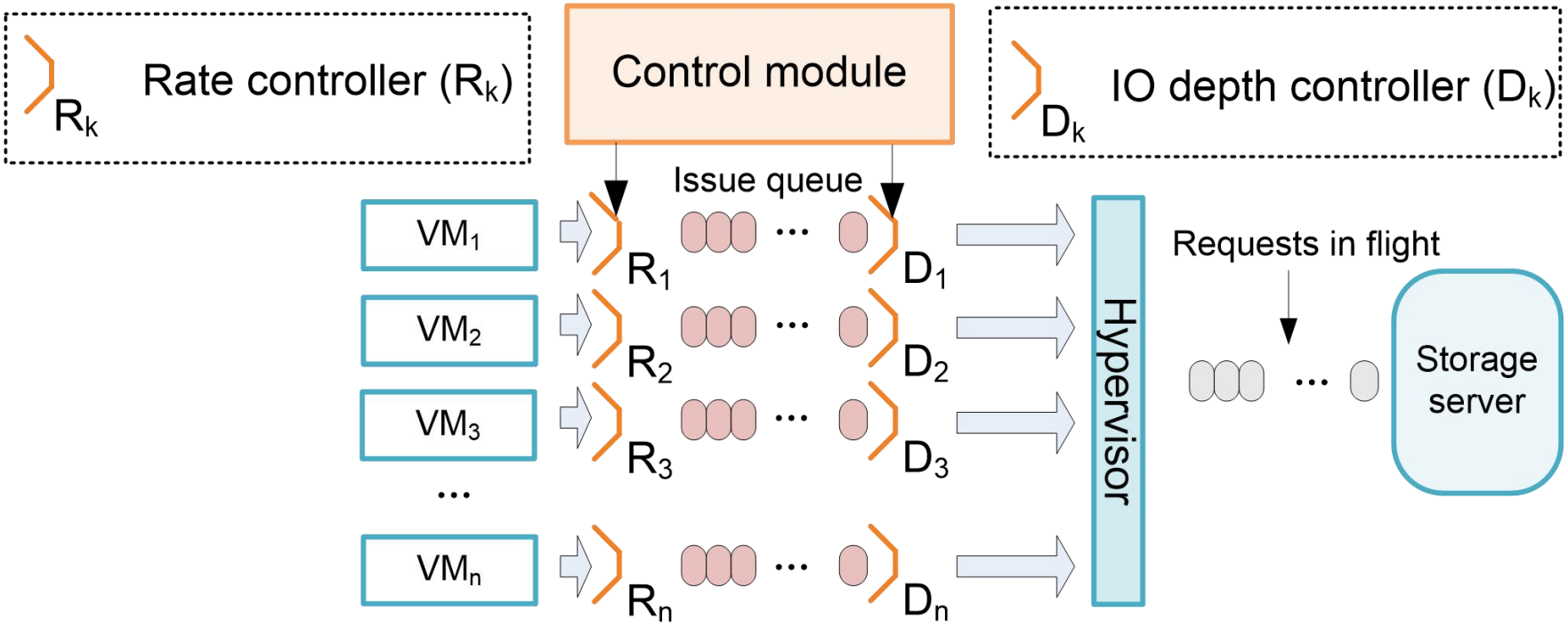
$$\zeta = 1 - \prod_{k=1}^{n} (1 - p_k)$$

虚拟机VM_k侦测未来I/O延时违例的概率：

$$p_k = \Pr(ST_k \geq H)$$

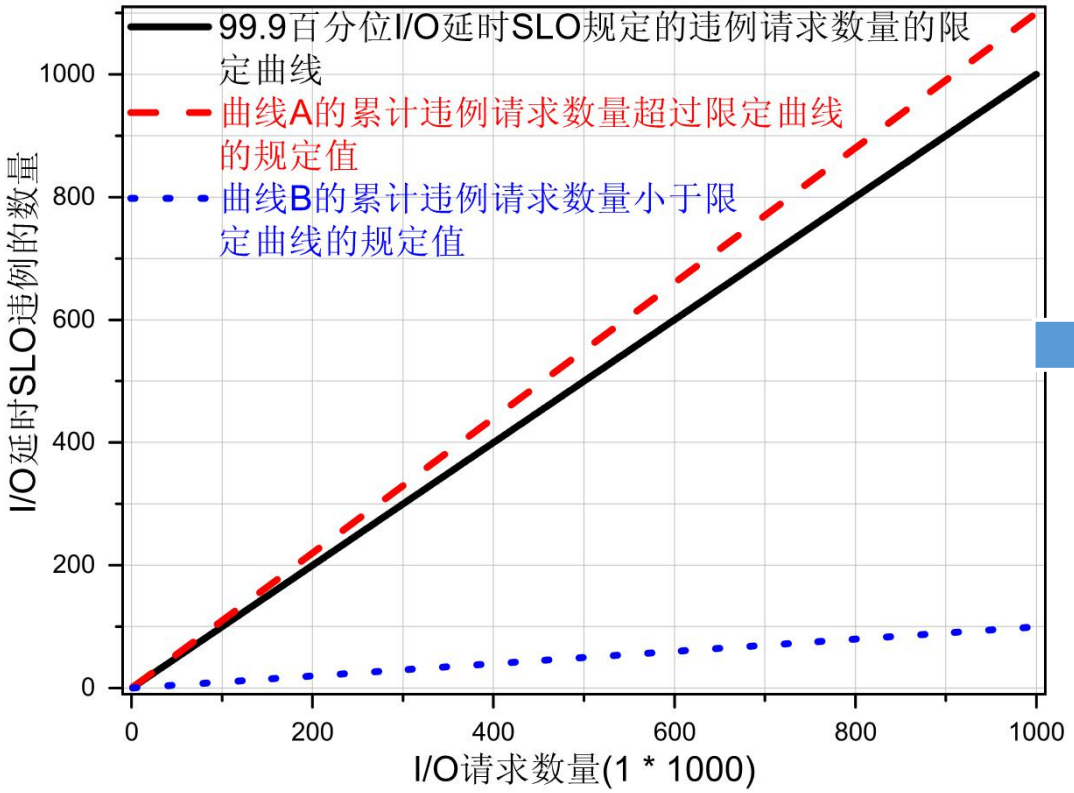
预测模块充分利用共享存储的虚拟机有I/O延时违例的时间局部性这一特征进行预测。

控制模块



控制模块通过动态调整IO并行度（Di）和虚拟机的IO到达率（Ri）对虚拟机级的既定百分位延时和I/O吞吐率进行控制。

决策模块



- 曲线A位于边界线之上, 导致99.9百分位延时SLO的违例。
- 曲线B 远在边界线之下, 这意味着很大的I/O吞吐率优化空间.
- 最理想的情况是曲线和边界线重合，这样I/O吞吐率将被最大化。

控制模块通过动态调整IO并行度（Di）和虚拟机的IO到达率（Ri）对虚拟机级的既定百分位延时和I/O吞吐率进行控制。

决策模块

实际的I/O延时违例数和上限的差值

$$e(k,t) = G(X_k,t) - M_k(t) \quad (6)$$

$$E(k,t) = \frac{e(k,t)}{D_k(t)} \quad (7)$$

使 e(k,t) 独立于D_k，因为它与负载I/O特征和 延时SLO有关

$$A(k,t) = Max(\sigma * \frac{e(k,t)}{G(X_k,t)}, 0) \quad (10)$$

$$P(t) = Min_{1 \le k \le n} A(k,t) + 1 \quad (11)$$

P(t) 是一个比例因子，它可以决定下一个时段所有聚合虚拟机I/O吞吐率目标提升的程度

如果百分位延时SLO能够得到满足，则 P(t) > 1

如果新更新的吞吐率目标是可行的, …

$$Th_k^G(t+1) = \begin{cases} P(t) * Th_k^G(t) & if\ \forall j, Th(j,t) \ge Th_k^G(t), \\ Max(\eta * Th_k^G(t), Th_k^{SLO}) & otherwise. \end{cases} \quad (12)$$

$$R_k(t+1) = Max(\frac{Th_k^G(t)}{Th_k(t)} * R_k(t), Th_k^G(t)) \quad (13)$$

如果百分位延时SLO不能得到满足， E(j,t) < 0

$$D_k(t+1) = \begin{cases} 1 & if\ \exists j, E(j,t) < 0, \\ Max(D_k(t)-1, 1) & else\ if\ Th_k(t) > Th_k^G(t), \\ D_k(t)+1 & else\ if\ E(k,t) \ge U, \\ D_k(t) & otherwise. \end{cases} \quad (14)$$

Th(k, i): 第i 时段VM_k 的实际吞吐率

$Th^G_k(t+1)$:第t+1 时段VM_k 的吞吐率目标

两维SLO约束的I/O优化

性能测试

测试平台介绍：

主机: 2台 PowerLeader PR2760T servers
*) 2 Intel Xeon E5620 quad-core 处理器,
*) 12GB 内存,
*) 40Gbps Mellanox MT26428 ConnectX VPI Infiniband网卡。

存储: 2类存储子系统
*) 16-disk (7200RPM, 250GB) RAID 0 磁盘阵列
*) Fusion-io ioScale 2, 825GB Multi Level Cell (MLC) SSD

虚拟化软件: Xen 4.2

测试目标:

*) 评测PSLO在任意百分位延时SLO约束下优化吞吐率和保障吞吐率分配的公平性的能力。
*) 评测PSLO在吞吐率SLO约束下优化任意百分位延时的能力。

性能测试

测试场景

MSN (VM数量: 25 VMs, I/O延时 SLO: 100ms，90% 服从率, I/O吞吐率 SLO: 3750 IOPS)

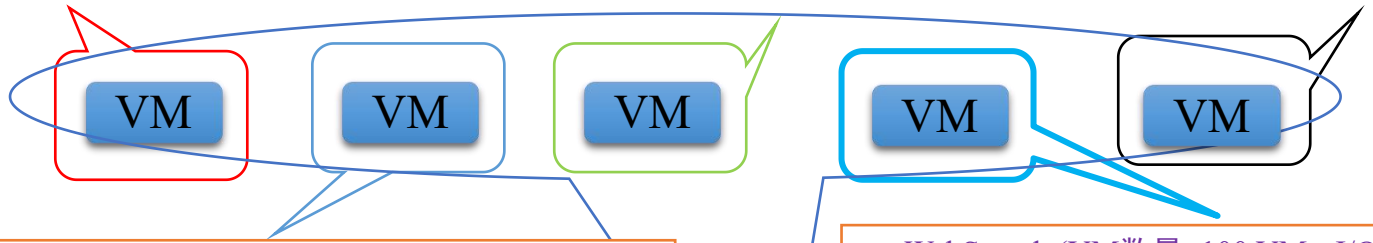
VM-level (I/O吞吐率SLO: 150 IOPS
99.6th 百分位I/O延时SLO: 100ms)

TPC-E (VM数量: 100 VMs, I/O延时 SLO: 200ms， 90% 服从率, I/O吞吐率 SLO: 100000 IOPS)

VM-level (I/O吞吐率: 100 IOPS
99.9th百分位I/O延时SLO: 200ms)

File copy (VM数量: 100 VMs, 无I/O延时 SLO, I/O吞吐率 SLO: 15000 IOPS)

VM-level (I/O吞吐率SLO: 150 IOPS
无既定百分位I/O延时SLO)



Exchange (VM数量: 15 VMs, I/O延时 SLO: 100ms，90% 服从率, I/O吞吐率 SLO: 2250 IOPS)

VM-level (I/O吞吐率 SLO: 150 IOPS
99.3th -ile百分位I/O延时SLO: 100ms)

WebSearch (VM数量: 100 VMs, I/O延时 SLO: 200ms，90% 服从率, I/O吞吐率 SLO: 10000 IOPS)

VM-level (I/O吞吐率SLO: 100 IOPS
99.9th百分位I/O延时Latency SLO of 200ms)

共享存储构架

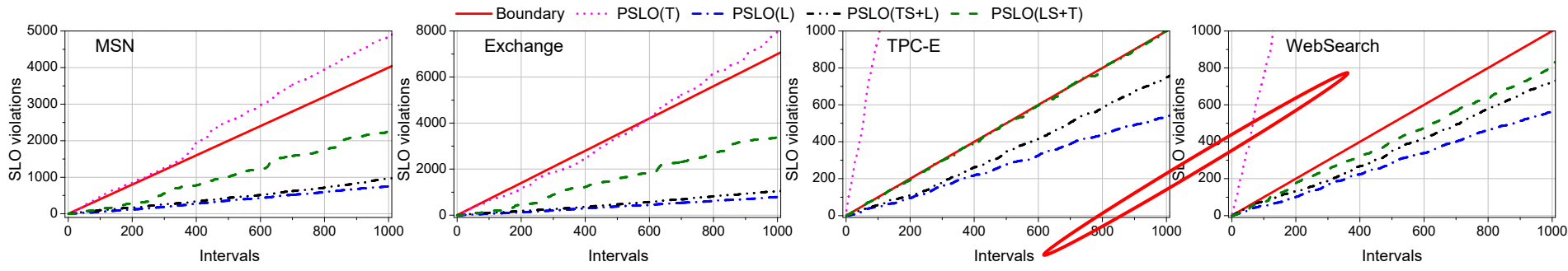
性能测试

SLO策略

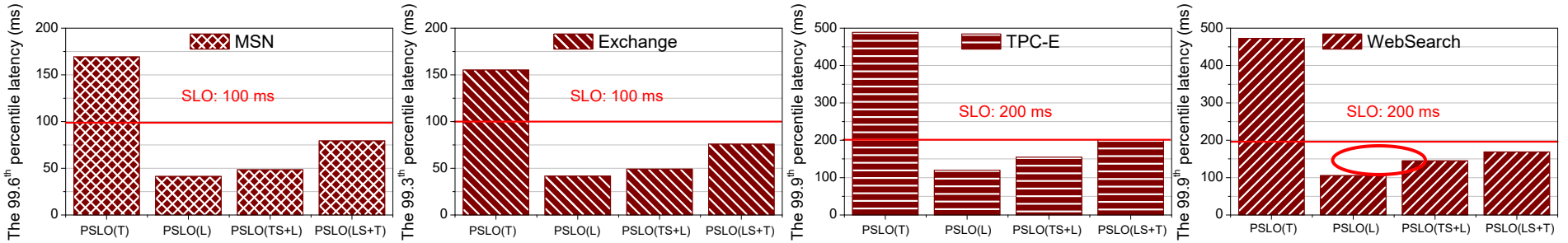
策略	保障吞吐率SLO	保障既定百分位延时SLO	吞吐率分配公平性	吞吐率优化	延时优化
PSLO(TS+L)	Y	Y	Y	N	Y
PSLO(LS+T)	Y	Y	Y	Y	N
PSLO(L)	N	Y	N	N	Y
PSLO(T)	Y	N	Y	N	N
No PSLO	N	N	N	N	N

性能测试

不同SLO约束下的性能优化



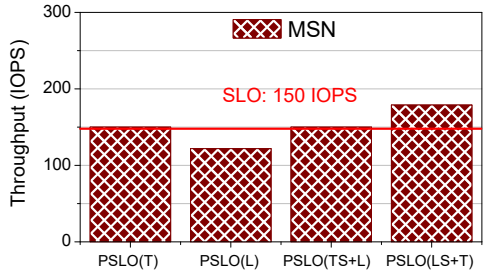
The 99.6 th-ile **The 99.3 th-ile** **The 99.9 th-ile** **The 99.9 th-ile**

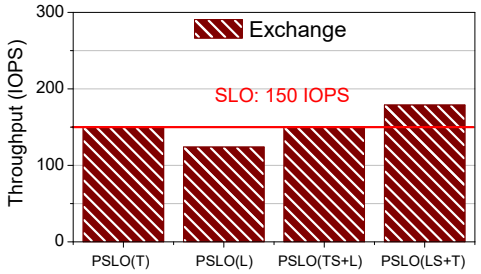


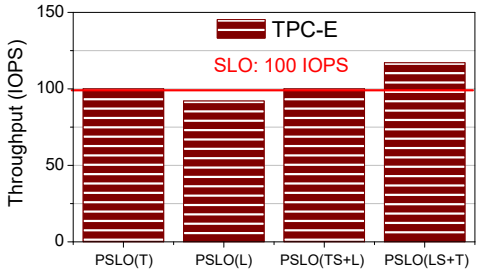
PSLO(LS+T)控制下, 既定百分位延时(200.4 ms)几乎和SLO 目标 (200 ms)一致。

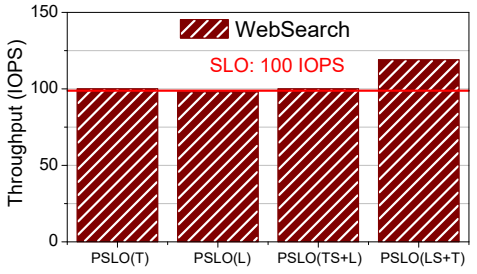
性能测试

不同SLO约束下的性能优化

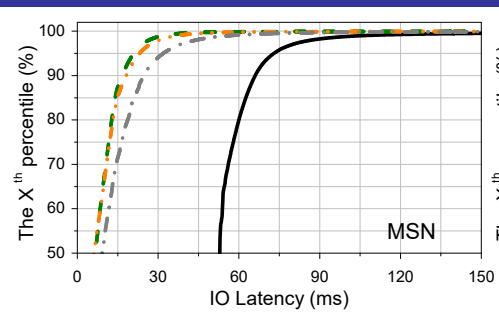


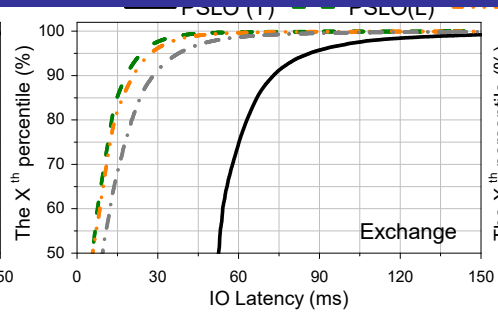


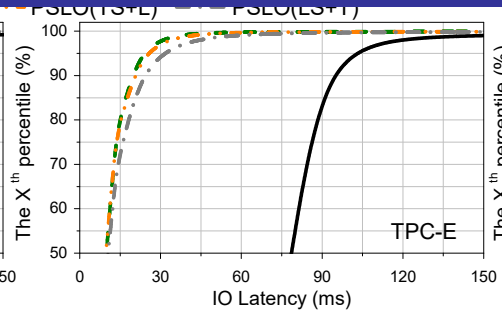


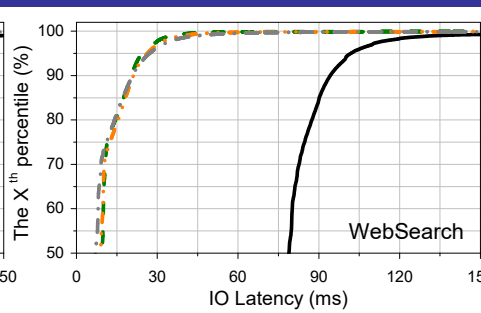


在PSLO(LS+T)控制下，四台虚拟机吞吐率分别相对于其吞吐率SLO提升19%, 19%, 18% 和 19%









PSLO(TS+L) 则能在准确保障吞吐率SLO的前提下优化既定百分位延时

本讲小结

- 提出聚合虚拟机环境中面向高存储资源利用率的尾延时SLO精确保障方案，该方案可在精确保障尾延时SLO的前提下优化I/O资源分配。
- 提出聚合虚拟机环境下既定百分位I/O延时和吞吐率SLO精确保障以及基于SLO的性能优化方案，该方案可以支持一台主机上运行的多个虚拟机具有完全不同百分位的I/O延时SLO和I/O吞吐率SLO，并实现基于上述两维SLO约束的I/O性能优化。

总结

- 数据中心虚拟化平台实践
- 虚拟化环境中存储系统面临什么问题
- 怎样进行服务质量精确保障
- 怎样在此基础上进行多维保障及优化