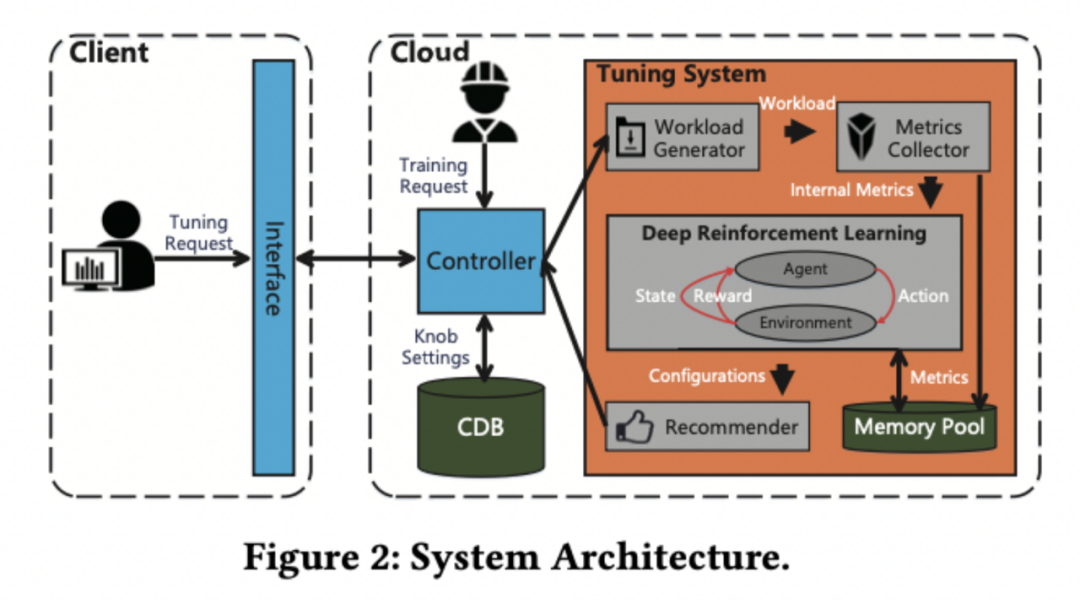
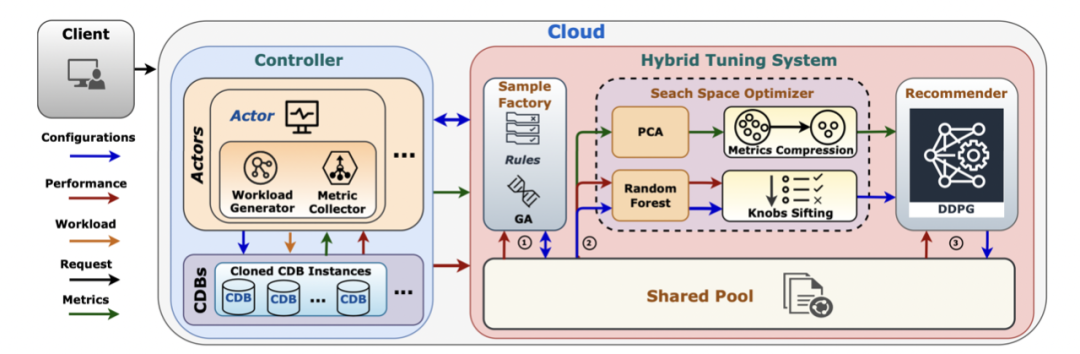
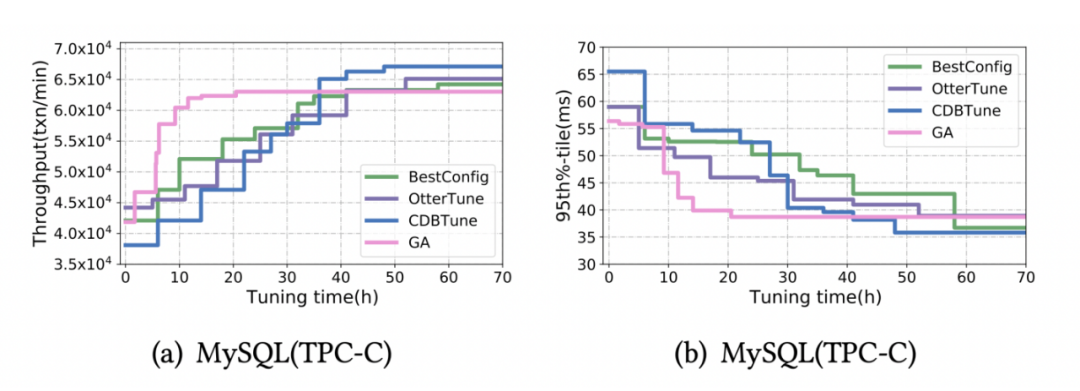
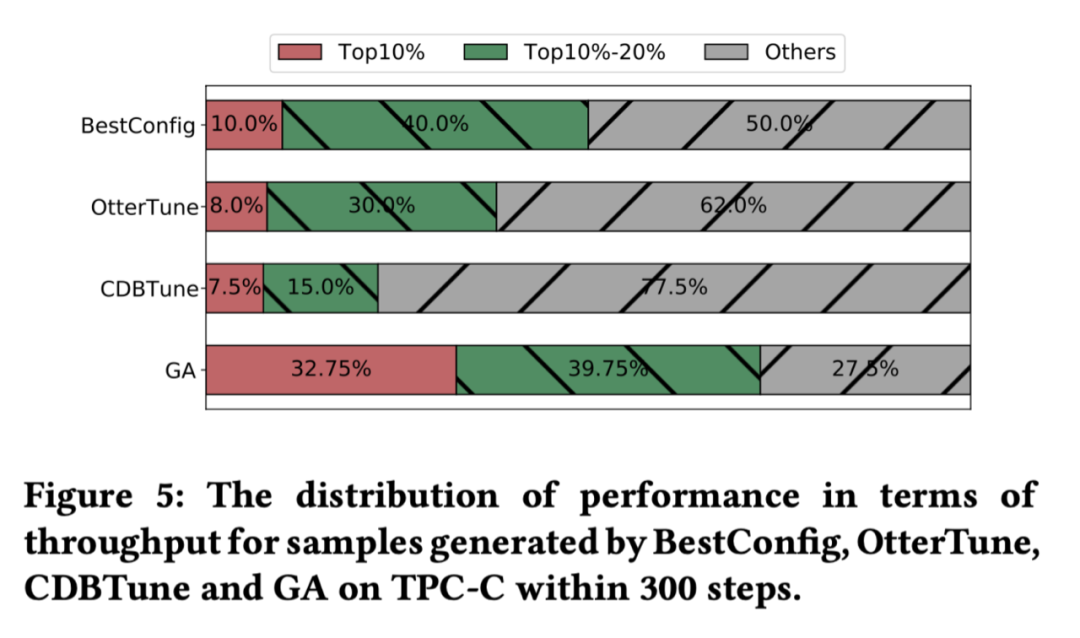
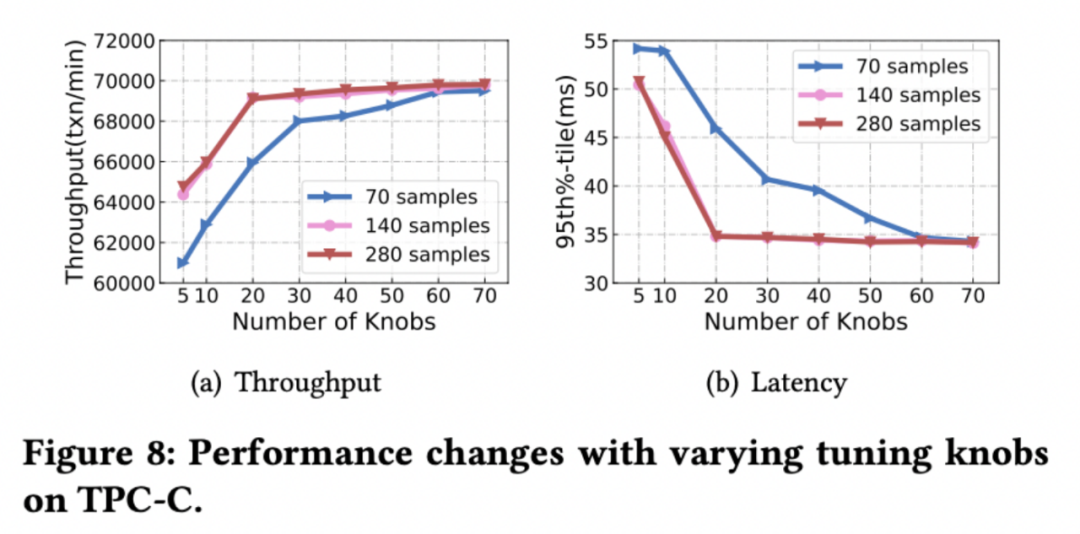
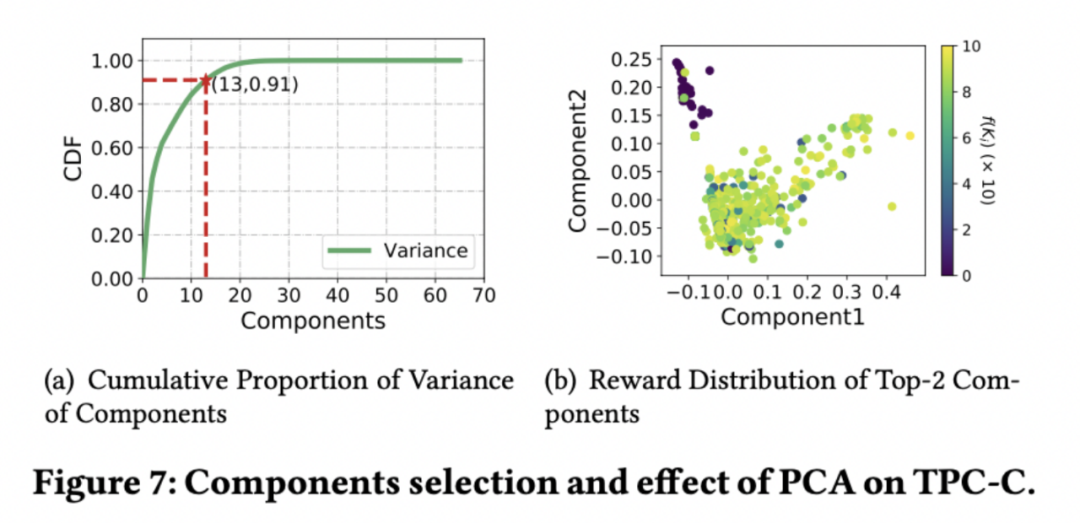
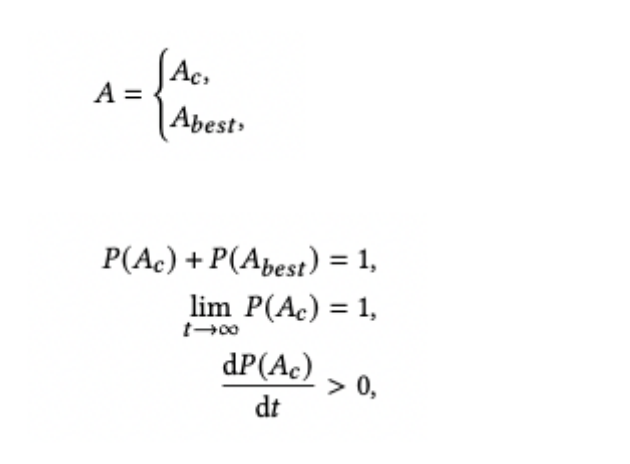
面向个性化需求的在线云数据库混合调优系统 | SIGMOD 2022入选论文解读

[腾讯云数据库](javascript:void(0);) 2022-06-30 21:54 发表于北京

SIGMOD 数据管理国际会议是数据库领域具有最高学术地位的国际性会议，位列数据库方向顶级会议之首。近日，腾讯云数据库团队的最新研究成果入选 SIGMOD 2022 Research Full Paper（研究类长文），入选论文题目为“HUNTER: An Online Cloud Database Hybrid Tuning System for Personalized Requirements”。标志着腾讯云数据库团队在数据库AI智能化上取得进一步突破，实现性能领先。  
数据库参数自动调优在学术界和工业界都已有较多研究，但现有的方法在缺少历史数据时或是面对新负载进行参数调优时，往往面临着调优时间过长的问题(可达到数天)。在此篇论文中，团队提出了混合调优系统Hunter，即改进后的 CDBTune+，主要解决了⼀个问题：**如何在保证调优效果的前提下显著减少调优时间。经实验调优效果明显：随着并发度提升实现调优时间准线性降低，在单并发度场景下调优时间只需17小时，在20并发度场景下调优时间缩短至2小时。**

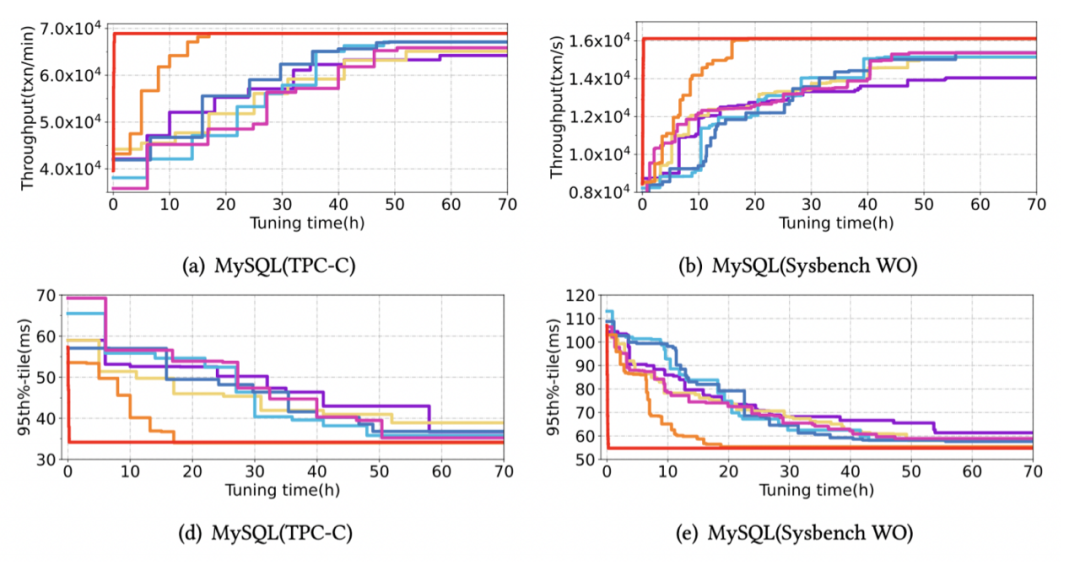
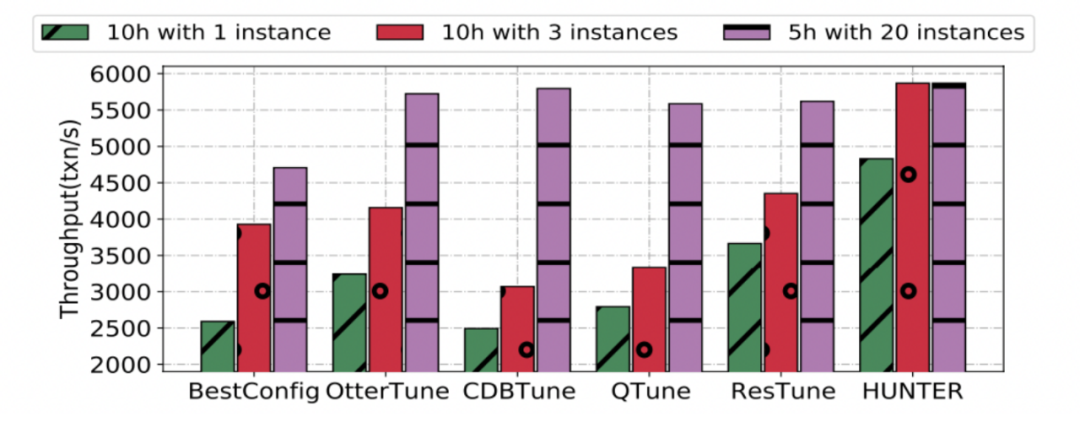
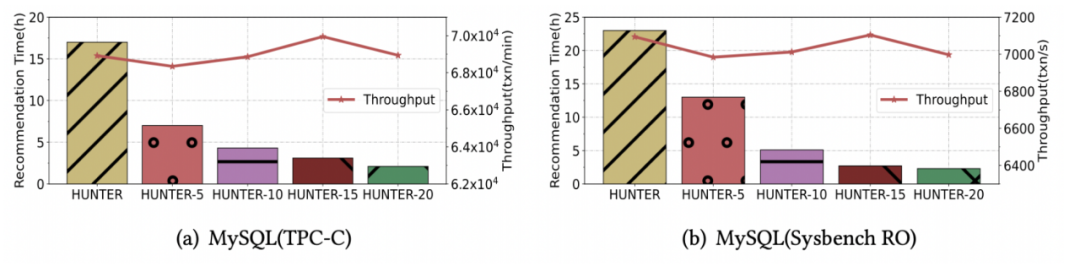
**工作原理**

**（技术原理解析）**

这是CDB/CynosDB数据库团队**第三次**研究成果论文被SIGMOD收录。继2019年数据库团队首度提出基于深度强化学习(DRL)的端到端云数据库参数调优系统CDBTune，该研究论文“An End-to-End Automatic Cloud Database Tuning System Using Deep Reinforcement Learning”入选SIGMOD 2019 Research Full Paper（研究类长文）。  
  
虽然CDBTune 在调参效果上已经达到了⼀个相当高的水平，但我们也发现，CDBTune 需要较长的调优时间才能通过自我学习达到较高的性能。  
对此，本次收录论文中提出改进的 CDBTune+，能够在保证调优效果的前提下极大地缩减调优耗时。  
**改进的混合调优系统CDBTune+，主要包含样本生成、搜索空间优化、深度推荐三个阶段。**样本生成阶段利用遗传算法进行初期调优，快速获取高质量样本；搜索空间优化阶段利用上⼀阶段的样本信息减小解空间，减少学习成本；深度推荐阶段利用之前阶段的信息进行维度优化和强化学习预训练，保证调优效果的同时显著减少调优时间。  
为了进⼀步对调优过程进行加速，我们充分利用CDB 的克隆技术，采用多台数据库实例实现并行化， 令整个调优时间更进⼀步地减少。  
  
**样本生成**  
如下图所示，由于基于学习的调优方法在训练初期都有着调优效果差、收敛速度慢等问题（我们称之为冷启动问题）。  
我们认为这些方法面临冷启动问题主要是因为：  
1、样本数量少质量差，网络难以快速学到正确的探索方向。2、搜索空间大，网络结构复杂，学习速度缓慢。  
为了缓解上述问题，**我们采用收敛速度更快的启发式方法（如：遗传算法(GA)）进行初期的调优，以此快速获得高质量的样本**。  
如图 5 所示，不同方法进行 300 次的参数推荐，图中是这 300 次参数所对应的数据库性能分布。可以见得，相较于其他的方法，GA 能够收集到更多的高性能参数。  
虽然有着更快的学习速度，但是 GA 却可能更容易收敛到次优解，如图 6 所示。  
  
启发式方法虽有着较快的收敛速度，但是却容易收敛到局部最优，导致最终调优效果不佳。  
而基于学习的方法却在较长的调优时间后可以得到较高的性能，但是却需要较长的训练时间，速度较慢。我们将两种方法结合，即加快了调优速度，也确保了参数质量。  
**搜索空间优化**  
  
单纯地将两者拼接难以有⼀定的性能提升（节约约 20%的时间），但是我们期望更多。  
利用样本生成阶段可以获得较多高质量的样本，但是却没有将其效果充分发挥。我们利用PCA 进行状态空间降维，Random Forests 进行参数重要性排序。  
PCA 是⼀种常用的降维方法，可将高维数据降为低维数据的同时保留大部分信息。我们采用累计方差贡献率来衡量信息的保留度，⼀般来说，当累计方差贡献率 > 90%时即可认为信息得到了完全的保留。  
我们选择贡献率最大的两个成分，并以此作为 x、y 轴描点，以其对应的数据库性能作为点的颜色（颜色越深性能越低），可以看出，低性能的点可以被两个成分较为明显的区分开来，由此可见，**PCA 能够帮助 DRL 更好地学习**。  
随机森林可以被用来计算特征的重要性，我们以数据库参数为输入，对应的数据库性能为输出训练随机森林模型，然后计算各个数据库参数的重要性，并进行排序。采用不同数量的 Top 参数进行参数调优可以看到数据库最优性能的变化，在⼀定数量的样本保证下，TPC-C 负载调整 20 个参数即可达到较高的性能。  
  
**深度推荐**  
经历样本生成和搜索空间优化后，我们在深度推荐阶段**采用深度强化学习（DRL）来进行参数推荐**。  
首先，搜索空间优化的结果会对 DRL 的网络进行优化，减少其输入输出的维度，简化网络结构。  
其次，样本生成阶段的样本将加入DRL 的经验池中，由 DRL 进行⼀定程度的预训练。最后，DRL 将基于改进后的探索策略进行参数推荐。  
DRL 的基本结构与 CDBTune 类似，为了充分利用高质量的历史数据，我们修改了其探索策略。动作 （数据库参数配置）有⼀定概率在历史最优参数附近探索，具体的计算方法如下图所示。 Ac 表示 DRL 的结果，Abest 表示历史最优，初始情况下 Ac 的概率为 0.3。

**调优效果性能分析**

**效果分析**

为了测试不同调优方法从零开始进行参数调优的效果，我们在不同负载下进行了测试。在测试中，所有方法都没有任何的预训练。其中 HUNTER-20 表示以 20 个实例进行并发调优的 HUNTER。  
  
如下图所示，虽然只有我们的方法提供了并发功能，但是并发加速本身是通用的，因此，我们在真实负载下对不同方法做了进⼀步测试。虽然大部分方法借助较长的调优时间可以获得足够高的性能，但是，在相同的代价情况下 （时间\*实例数），HUNTER 的表现是最好的。  
  
下图展示了 HUNTER-N 达到串行所能找到的最优性能的调优耗时，可见调优速度的效果，**随着并发度增加，调优时间显著缩短**。  
  
**不足**  
对于 DBA 来说，负载越简单所需的调优时间应该会越短，但是自动调优方法却没有这样的特质，如我们上述的实验图所示，有些时候，简单负载可能需要更多的时间来获得更高的性能。更重要的在于，我们目前难以快速地判断性能是否达到了“最优”，这导致我们花费了额外的时间来观察调优系统是否能令数据库性能再得到提高。

**目前**

通过技术解读和效果分析，我们可以看出改进后的Hunter大幅提升调优效果，同时体现出论文对实际数据库问题的落地可能性很高，具有指导方法意义。

在接下来的研究中，我们希望结合专家经验来解决上文提到的问题，提高参数调优的可解释性并更进一步压缩调优时间，同时也希望找到一种估计最优性能的方法，从而减少额外的调优时间。  
CDBTune+旨在降低数据库参数调优的复杂度，实现参数调优零运维，是腾讯云数据库AI智能化变革的再一次跨越和实现。智能调优一期已经在腾讯云MySQL产品上线，后续会在更多腾讯云数据库产品上应用，为学术及工业界带来更多贡献和服务。

﹀

﹀

﹀

**-- 更多精彩 --**

[](http://mp.weixin.qq.com/s?__biz=Mzg4NjA4NTAzNQ==&mid=2247503801&idx=1&sn=192e309e6da98ac8a6dd7897da34f30a&chksm=cf9d8b7df8ea026b4bb8164d9825626e34b6a0d6b86ff273e522a62c0838853896e18ed6fc4e&scene=21#wechat_redirect)

必看！重点高校教授带你读数据库顶会论文

[](http://mp.weixin.qq.com/s?__biz=Mzg4NjA4NTAzNQ==&mid=2247500850&idx=1&sn=c3f32ccd215ff1cff5c3aa29c121519a&chksm=cf9d80f6f8ea09e09b0cd0652b200ef29c6127ec718ebeb1d3435d13dc967d7f344eb980d730&scene=21#wechat_redirect)

解放双手，数据库智能调参CDBTune等你来测

**↓↓点击阅读原文，了解更多优惠**

[阅读原文](javascript:;)

阅读 466



腾讯云数据库

分享收藏1

赞

​



写下你的留言



复制搜一搜分享收藏划线