Characteristics of Co-Allocated Online Services and Batch Jobs in Internet Data Centers: A Case Study From Alibaba Cloud

摘要

本文探讨了来自阿里云包含1.3k服务器的生产集群中共同分配的在线服务和批处理作业的各种特征。

研究发现：1）对于具有多个任务和实例的批处理作业，当它们的第一个和唯一一个实例失败时，经过很长的时间间隔后，有50.8％失败的任务等待并暂停。2）对于在线服务作业，可根据其请求的CPU，内存和磁盘资源将它们分为25类。3）服务器可按CPU利用率，内存利用率及其相关性分为七个组。4）实例的MTBF（平均故障间隔时间）在[400，800]秒之间，而99%的平均完成时间为1003秒。

本文还比较了作业和服务器的累积分配功能，并解释了它们之间工作负载分配的差异和机会。

引言

本文尝试回答以下问题：

1）如何根据作业的资源需求和服务器的资源可用性将作业调度到适当的服务器。

2）如何以一定百分比降低在线服务的延迟，并提高批处理作业在同一群集上混合时的吞吐量，以及如何减轻此群集上的资源争用并重新调整工作负荷计划。

3）在硬件和软件故障的情况下，如何根据作业服务器的亲和力提供容错调度。

4）如何基于工作负载特征模拟真实的Internet数据中心（IDC），以进行更好的调度程序设计。

数据集说明

阿里巴巴跟踪数据ClusterData201708包含24小时内生产群集的群集信息，并包含1313台同时运行在线服务和批处理作业的计算机。表1中列出了本文中使用的符号和术语。

批作业的工作负载特征

A实例完成

实例是批处理作业调度的最小单位。图1给出了所有批处理实例的完成时间的CDF图表，表2给出了它们的百分位数。

B作业和任务分配

所有批处理作业都被划分为多个任务，每个任务均被划分为多个实例。

为了获得任务和作业的特征，图（2.a）给出了它们的编号。结果显示，大多数作业的任务少于40个。图（2.b）显示jobID为4000至5000和7000至9000的作业具有相似的实例编号。

图3显示了它们的CDF图，表3显示了它们的百分位。结果显示，80%的任务号和实例号分别是11和940。

图4和图5给出作业完成时间，任务编号和实例编号。数据显示，当实例数量在1000之后增加时，作业完成时间显著减少。

图6给出了所有作业的完成时间和完成时间小于1000s的作业。数据显示，ID约为10000的作业的完成时间比其他作业要少得多。

C资源利用

图7和图8给出了内存和CPU利用率。数据显示，内存利用率要比CPU高；内存利用率在24小时内会定期变化。

基于此，图7（b）给出了存储器利用率的拟合曲线，拟合公式为y=35\*（0.00129\*x-50.0828）+45，其中y是内存利用率，x是时间。拟合正弦曲线大小为35（％），周期为4870秒。

D作业故障分析

在跟踪数据中，有15739375个实例（不包括重新运行的实例）和209168个失败实例，它们的故障率均为1.32％。

表4和图9列出了故障统计信息。数据显示，在某些机器上执行的实例的故障率高于其他机器；某些任务和作业的失败实例数高于其他实例。

表5列出了故障实例编号最高的前15个计算机，图10给出了被丢弃的1121实例在计算机上的分布。数据显示，大多数机器丢弃了0-5个失败的实例，但是有些机器丢弃了10个以上的实例。

图11给出了失败的作业随时间的分布。数据显示，时间戳60352之后不再有故障；在时间戳之前，作业ID为[4000，5000]和[7000，9000]的作业不会失败。

在线服务工作负载特征

A实例聚类

本研究在资源分配上使用K-means聚类，当选择25作为聚类数时，聚类性能最佳，如图12所示。使用较低的聚类得分，将资源分配配置分为25种，聚类结果如表6所示。数据显示，大多数实例具有较低的CPU使用率和较高的内存使用率。

B资源利用

图13和图14列出了所有在线实例的CPU和内存利用率的总和。数据显示，内存使用曲线比CPU使用情况更稳定；在时间戳50100和54300上，CPU和内存利用率均出现急剧下降。

节点用途分析和表征

A机器集群

图15列出了数据中心中所有服务器的总CPU和内存利用率。数据显示，整个集群的CPU利用率在13％到40％之间，内存利用率在42％到65％之间。

表7列出了整个群集中所有计算机的群集结果。数据显示，类型A的计算机是群集中的大多数，它们平均具有27.5％的CPU利用率和50.67％的内存利用率。

图16列出了每种类型机器的CPU和内存利用率，包括CPU使用率和内存使用率。根据其不同的曲线形状分类，总共七个类别。数据显示，类型A的比例最大，达到总数的84.77％，其曲线也是低下降的近似直线，其他类别在某些时间或所有时间都有一些异常曲线出现，计算出的总体图与图（16.a）非常相似； B型机器在时间戳52000之后变为空闲；C型计算机在CPU和内存利用率上具有相似的模式，尤其是在时间戳为70000的突发并且该突发持续约10000秒时；在此跟踪数据的采样期间，十七台D型计算机似乎处于空闲状态；类型E，F和G的机器太少，无法表示任何模式。

图17中给出了七种类型机器的CPU使用率和内存使用率的差异。

B.CPU和内存利用率的相关性分析

为了量化在线服务作业和批处理作业的干扰，表8将计算机分类为3种类型，图18提供了每台计算机的CPU和内存利用率的相关性，图19说明了这三个机器子集的CPU和内存利用率。由此可以推断在线作业和批处理作业调度的共同分配的相关性，并估计共同分配后它们的CPU和内存利用率。

图20绘制了每台机器上CPU和内存利用率的相关系数。

C机器的MTBF

图21计算了所有机器上实例的MTBF（平均故障间隔时间）。结果显示，实例的MTBF在400秒到800秒之间。

结论

在线服务和批处理作业的混合会导致调度复杂性以及在线服务和批处理作业之间的干扰。此外，对在线服务的严格延迟控制限制了在线服务和批处理作业之间的资源复用。

本文分析了阿里云中生产集群中共同分配的在线服务和批处理作业的各种特征。本文对批处理实例的完成时间，资源利用率，故障分布，资源之间的关联和干扰以及机器运行特性进行了详细分析。