Internet数据中心中共同分配的在线服务和批处理作业的特征：以阿里云为例

抽象

为了降低电力和能源成本，巨型云提供商现在将在线和批处理作业混合在同一群集上。尽管此类作业的共同分配可以提高机器利用率，但在服务质量，容错和故障恢复方面，特别是对于延迟关键的在线服务而言，它对数据中心调度程序和工作负载分配提出了挑战。在本文中，我们探讨了来自阿里云中包含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。从跟踪数据中，我们发现以下内容：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。这将浪费大量时间和资源，因为其余实例正在运行，无法成功终止。 2）对于在线服务作业，根据其请求的CPU，内存和磁盘资源将它们分为25类。这样的群集可以帮助将在线服务作业与批处理作业同时分配。 3）服务器按CPU利用率，内存利用率及其相关性分为七个组。在CPU和内存利用率之间具有很强相关性的机器为作业协同分配和资源利用率估算提供了机会。 4）实例的MTBF（平均故障间隔时间）在[400，800]秒之间，而第99个百分位数的平均完成时间为1003秒。我们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。本文提供的研究结果和见解可以帮助社区和数据中心运营商更好地了解工作负载特征，提高资源利用率和故障恢复设计。

索引条款 共同分配的作业，工作负载表征，在线服务，批处理作业，数据中心，调度。

一，引言

功耗不仅成为云服务器的重要问题，而且也成为电池供电的设备的主要问题[1]，[2]。在过去的十年中，数据中心的用电量显着增加，这主要是由于数据中心的数量和密度都呈爆炸性增长。据估计，2014年美国数据中心的耗电量约为700亿千瓦时，占美国总能耗的2％。从2010年到2014年，数据中心的总能耗增长了4％，而从2005年到2010年，则增长了24％[3]。由于商业服务器的能源效率提高以及数据中心的节能行动，能源消耗的增长速度正在降低。特别是，提高能源效率是抑制数据中心行业能源消耗增长率的关键因素。如果能源效率的提高停滞在2010年的水平，据估计，数据中心将比2014年多消耗400亿千瓦时的能量[4]。除了提高硬件的能源效率外，还提出了各种节能的调度方法来节省数据中心的功耗和能耗[5] \_ [8]。

数据中心工作量的变化导致资源利用率的下降，这为资源复用提供了机会[9]，[10]。例如，当前的巨型云服务提供商在同一群集上共同分配在线服务和批处理作业，以提高服务器利用率并降低能源成本。因此，工作负载表征有助于了解工作负载模式并在云数据中心中设计更好的作业调度策略[11] \_ [13]。但是，在线服务和批处理作业的合并也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，这有时会降低系统性能。

此外，随着通信和传感器技术的发展，无线传感器已部署在智能城市，智能家居，自动驾驶汽车和工业环境中[14]。这些智能传感器能够提供具有多种要求的复杂服务，包括数据聚合和分析[15]。在新兴的边缘计算范例[16]中，可以在传感器节点中执行数据分析以节省能耗或隐私保护[17] \_ [20]。还有一些基于云的平台可为边缘设备提供数据分析，例如AWS Greengrass [21]和Microsoft Azure IoT Edge [22]。由于服务器和进来的工作负载的地理分布，云服务提供商倾向于采用过度提供策略来响应间歇性突发工作负载，以确保不同工作的服务质量。因此，精心设计的数据中心在线服务和批处理作业的混合不仅可以提供进一步的服务器整合，还可以确保租户的服务质量保证以及减少数据中心的能耗。

在本文中，我们分析了具有在线服务和批处理作业混合分布的数据中心的阿里云跟踪数据[23]。它是从具有1313台服务器的数据中心收集的24小时跟踪数据，包括资源利用率和作业状态。了解工作负载的特性对于工作负载的放置和调度以及服务供应的质量至关重要。我们尝试回答以下问题：

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。

本文的其余部分安排如下。在第二部分中，我们描述了跟踪数据，并给出了本文中使用的一些符号和术语。在第三部分中，我们分析了批处理作业，并描述了实例完成时间，作业和任务分配，资源利用率以及作业失败。第四节介绍了在线服务工作的分析。在第五部分中，我们提供了有关服务器节点的分析。我们在第六节总结了相关工作，并在第七节总结了本文。

二。分析数据集

A.阿里巴巴追踪数据

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。向公众公开这些数据可以帮助解决大型IDC面临的在线服务和批处理作业被共同分配的挑战。此跟踪数据的表征可以为在线服务和批处理作业计划程序的合作提供有用的见解。它还可以帮助平衡在线服务和批处理作业之间的资源分配，以平衡批处理作业的吞吐量，同时保持可接受的服务质量并快速恢复在线服务的故障。

B.度量符号和条款

为了方便起见，我们在表1中列出了本文中使用的符号和术语。

三，批作业的工作负载特征

A.实例完成

通常，一个批处理作业被分为多个任务，每个任务执行不同的业务逻辑。由于数据相关性，属于一个作业的任务由有向无环图（DAG）组成（尽管该数据集确实具有DAG相关性，但该数据集中没有发布DAG信息）。实例是批处理作业调度的最小单位。对于批处理，任务中的所有实例在具有相同资源请求但输入数据不同的情况下执行相同的应用程序代码。

我们在图1中给出了所有批处理实例的完成时间的CDF图表，在表2中给出了它们的百分位数。实例完成时间的第80、90和99％分别为132、260和1067秒。

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，我们在图（2.a）中给出它们的编号。我们可以看到，大多数作业的任务少于40个。jobID为4000至5000和7000至9000的作业具有相似的任务编号，可能的原因是这些作业相似。我们还可以在图（2.b）中观察到jobID为4000至5000和7000至9000的作业具有相似的实例编号。

我们在图3中显示了它们的CDF图，在表3中显示了它们的百分位。

第80个百分位的任务号和实例号分别是11和940。

在实际的作业调度方案中，一个实际的问题是确定作业完成时间是否与其任务和实例的划分（即作业分区粒度）相关。

我们在图4和图5中给出作业完成时间，任务编号和实例编号。当实例数量在1000之后增加时，作业完成时间显着减少。从图4中可以看出，大多数作业的实例数量少于1000，并且其运行时间是安静的多样性，与图4和图5相比，大多数任务编号和实例编号均小于1000，根据数据分析，我们知道99.2％的作业实例编号小于1000，而99.4％的作业任务编号小于1000。

我们列出了所有作业和完成作业的完成时间图6中小于1000秒的时间。

从图6中，我们可以看到ID约为10000的作业的完成时间比其他作业要少得多。对这些作业的调查可能有助于设计更好的作业调度策略，以缩短作业的完成时间。

C.资源利用

资源复用和作业协同分配可以显着降低能耗并提高资源利用率。因此，资源利用分析对于这种复用和共分配的质量评估至关重要。我们在图7和图8中给出了内存和CPU利用率。

从图7中我们可以看到，内存利用率要比CPU高。此外，内存利用率在24小时内会定期变化。因此，作业调度可以适应该特性以平滑内存使用。我们在图7（b）中给出了存储器利用率的\_tting曲线。量化为y D 35\_ .0：00129\_x 50：828 / C 45其中y是内存利用率，x是时间。

\_tting正弦曲线的大小为35（％），其周期为4870秒。

D.作业故障分析

健壮的作业调度策略对于数据中心运营中的业务一致性和质量服务提供非常重要。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％。我们在表4和图9中列出了故障统计信息。

从图9中，我们观察到在某些机器上执行的实例的故障率高于其他机器。同样，某些任务和作业的失败实例数高于其他实例。对于具有多个任务和实例的批处理作业，我们发现在1126个失败的任务中，有572个失败的任务（占失败任务总数的50.8％）在很长的时间间隔后等待并停止，这是它们的第一个任务和唯一一个失败实例发生。这将浪费大量时间和资源，因为其余实例正在运行，无法成功终止。这也表明作业调度程序应更智能，更快速地执行操作，以减少现有实例发生故障后的无用执行。

我们在表5中列出了故障实例编号最高的前15个计算机。这些计算机可能与硬件有关的问题或软件配置与在其上运行的实例不兼容。因此，应首先对这些机器进行检查以进行故障诊断。在表5中，删除的实例总数为1121。这1121个实例可能是系统管理员有意手动删除的，或者在实例过时后由调度程序自动删除的，或由于调度程序故障而删除的。扫描这些丢弃的实例可以提供有关作业调度和缓解这些丢弃的实例的见解。

在以下各节中，将服务器节点进行分组和分类，以更好地了解整个数据中心的操作模式。分类标准之一是其上失败实例的数量。

在所有实例的第一个执行期间，在148899个失败的实例中，在第二次执行期间删除了1121个实例。最终，包含这1121个实例的作业失败。由于这1121个丢弃的实例导致其父任务和作业失败，因此我们在图10中给出了这些1121丢弃的实例在计算机上的分布。

从图10中，我们观察到大多数机器丢弃了0-5个失败的实例，但是有些机器丢弃了10个以上的实例。

如果一个实例在执行期间失败，但是调度程序未通知另一个实例这种失败并继续执行，则其父任务或作业可能会在额外延长的执行时间之后最终失败。但是，这可能会浪费资源，降低性能或违反服务水平协议。为了导出实例作业的时间模式，我们在图11中给出了失败的作业随时间的分布。

从图11中我们可以看到：1）时间戳60352之后不再有故障； 2）在时间戳之前，作业ID为[4000，5000]和[7000，9000]的作业不会失败。这些作业在时间戳40000到60000之间失败，但是失败的作业具有连续的jobID和时间。这可能是由一些级联的作业失败引起的。

IV。在线服务工作负载特征

A.实例聚类

在阿里云数据中心中，在线服务和批处理作业被共同分配在同一集群上。在线服务由实例表示。在此跟踪数据中，有11101个实例，其中在跟踪数据采样之前创建了10980个实例，在跟踪数据采样期间仅创建了193个实例。

通常，将为实例分配内存，CPU和磁盘容量。 CPU分配由CPU内核数（例如1、2、4、8或16）表示。内存分配是归一化的值，例如0.002651、0.042430、0.053012、0.084819、0.127228、0.169637、0.254456、0.318000、0.999963、1.00002 ，1.000002或1.000010。磁盘分配也被标准化为3.17e-11、0.000142、0.034085、0.045446、0.056808、0.068170、0.085212和0.113617。因此，存在5\_13\_8 D 520个资源分配和107个CPU SET分配的组合。尽管有520种资源分配组合，但我们在资源分配上使用K-means聚类，因此，当我们选择25作为聚类数时，聚类性能最佳，如图12所示。 （请注意，在我们的实验中，使用较低的聚类得分，可以将资源分配配置分为25种以上。）聚类结果如表6所示。

表6中的聚类结果有助于估计传入的在线实例的资源分配并使其更好将这些联机实例调度到具有可用资源的适当群集节点。从表6中可以看出，大多数实例具有较低的CPU使用率和较高的内存使用率，因此，我们可以添加一些具有较高CPU使用率和较低内存使用率的任务。根据我的数据分析，82.58％的实例占用的磁盘容量不到20％。

B.资源利用

我们在图13和图14中列出了所有在线实例的CPU和内存利用率的总和。从图13和图14中，我们可以看到内存使用曲线比CPU使用情况更稳定。我们还观察到，在时间戳50100和54300上，CPU和内存利用率均出现两个急剧下降。这可能表示这两个位置服务中断或不可用。通过实时系统监视或日志流处理，我们可以快速诊断和定位此类在线服务异常。

五，节点用途分析和表征

A.机器集群

充分了解工作负载特征对于协调数据中心在线和批处理作业的工作调度至关重要。在前面的部分中，我们分析了工作特征。在本节中，我们介绍了集群中机器特征的分析。

我们在图15中列出了数据中心中所有服务器的摘要CPU和内存利用率。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。

在线服务和批处理作业在同一群集上的共同分配使资源复用和服务供应面临挑战。例如，在线服务对第99.9个百分点的工作有严格的延迟要求；因此，在专用计算机上共同分配作业之前，每台计算机的现有资源利用率是重要的参考。因此，我们在表7中列出了整个群集中所有计算机的群集结果。我们观察到，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。

我们在图16中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，下图显示了总共七个类别。表7示于图16。计算了七种类型，不同类别的比例，平均CPU和内存使用率，以及CPU和内存使用率的中位数和中位数。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。结合图形和表格，我们知道类型B和类型C节点的平均内存和CPU低于类型A的平均内存和CPU，但是方差高于类型A的方差。可以认为B型和C型节点上的任务模式更加复杂，并且需要一些更加专门的任务调度。在图16中，对于B型机器，它们在时间戳52000之后变为空闲。由于某些原因，可能计划对这154个B型机器进行预定维护或停用。如果它们只是闲置的并且可以处理传入的请求，那么让它们长时间闲置不是一个好主意，并且可以有意地关闭它们的电源。

对于C型计算机，它们在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。这21个C型计算机在工作负载爆发期间几乎使用100％的CPU。应该将它们上的作业迁移到其他计算机上，以使计算机保持相对较低的CPU利用率，以确保足够的服务质量。

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，这意味着这些计算机处于空闲状态或已停止运行。类型E，F和G的机器太少，无法表示任何模式。我们在图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，我们在表8中将计算机分类为3种类型，并在图18中提供了每台计算机的CPU和内存利用率的相关性。我们在图19中说明了这三个机器子集的CPU和内存利用率。在大多数计算机上，它们的CPU和内存利用率呈正相关。通过作业类型和实例分布，我们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。

我们在图20中绘制了每台机器上CPU和内存利用率的相关系数。对于CPU和内存利用率具有强正相关性的类型2机器，我们用机器ID标识这些机器，即机器88-127，机器261-296和机器830-906。它们可能是三个小型集群，包含所有1313台计算机的子集，这些计算机与它们的邻居计算机执行相同的作业。对于这些机器，如果计划了更多相同类型的作业，则可以轻松估算机器利用率。具体来说，如果我们将这些计算机调度到群集中新添加的计算机，我们还可以估计专用计算机的利用率。此外，如果我们将更多的计算或内存密集型作业打包到现有计算机上，则还可以根据CPU和内存利用率之间的相关性来估算以下资源利用率。因此，CPU和内存利用率的相关性可以帮助在线服务和数据中心中的批处理作业的共同分配和合并。

C.机器上的MTBF

如前一节所述，某些计算机上的实例比其他计算机（例如＃679，＃680，＃341和＃673）更容易出现故障。相反，有些机器上的实例永不失败，例如，机器＃167，＃478和＃1075。这些机器的比较可以提供有用的见解和指导，以诊断与硬件或软件有关的问题，从而避免实例故障。我们计算所有机器上实例的MTBF（平均故障间隔时间），如图21所示。由于不同的计算机具有不同的故障模式，因此我们没有在每台计算机上给出MTBF编号。或者，我们为大多数机器提供MTBF范围。在大多数计算机上，实例的MTBF在400秒到800秒之间。实例的MTBF在三个方面非常重要：

1）应该考虑确定是否值得将作业调度到这些计算机，尤其是对于那些需要高成功率或较长执行时间的作业。

2）它有助于估计在不久的将来何时以及有多少个作业会失败，或者帮助您将实例迁移到其他计算机上以避免失败。

3）如果计划的机器维护或停止运行，则它是很好的参考。

VI。相关工作

工作负载表征是数据中心操作员识别系统瓶颈并提供解决方案以优化性能的关键过程。单服务器工作负载特征已得到广泛研究。科尔特斯等。 [24]展示了来自Microsoft Azure的几种VM工作负载行为的详细特征。他们通过虚拟机的生命周期，部署大小和资源消耗来分析工作负载的关键特征。郭等。 [25]将工作负载分为两类：延迟敏感的交互式工作负载和耐延迟的批处理工作负载，他们提出了随机成本最小化算法（SCMA）。在具有可再生能源的时间敏感的，地理分布的数据中心[26]中，作者采用了绿色能源预测来计划数据中心中的混合批处理和服务工作。同样，一些研究人员表明，在加速性能和微体系结构特征方面，数据分析工作量明显不同[27]。 Chen等。 [28]和Chong等。 [29]认为工作负载表征对于优化数据中心的工作负载管理很重要，并且他们可以根据新出现的工作负载来动态地做出决策。 Wu等。 [30]提出了Dynamo-a数据中心范围的电源管理系统。 Fiandrino等。 [31]提出了一种新的度量框架，该框架能够评估云计算通信系统的性能和能效。为了在不降低工作负载性能的情况下限制峰值用电成本，Aksanli等人提出了建议。 [26]，Aksanli和Rosing [32]采用基于电池的峰值剃须方法，而Liu [33]为数据中心开发了两种算法。

服务器整合和VM迁移是提高数据中心资源利用率ef\_ciency的典型方法。 Ahmad等。 [34]分析和比较当前的VM迁移和服务器整合框架。 Varasteh和Goudarzi [35]调查了虚拟机迁移和服务器整合，用于将VM整合到PM的参数和算法方法。 Mastroianni等。 [36]提出了ecoCloud，一种用于在两个资源（即CPU和RAM）上整合VM的自组织和自适应方法，以限制VM迁移和服务器切换的次数。同时，其他人则专注于Xen和VMware虚拟化平台[37]。另外，由于合并不同的应用程序可能会导致性能下降，Chen等人。 [38]开发一种轻量级，非侵入性的方法来实现以应用程序为中心的性能目标，同时巩固同质和异构的应用程序。

数据中心在资源分配方面有大量工作。 Sun等。 [39]，介绍了数据中心不同种类的资源管理机制。 Tan等。 [40]和Mazumdar和Kumar [41]提出了各种方法来分析资源使用情况并为资源使用模式建模。另外，为了维持较高的资源利用率，已经提出了针对CPU和内存的新资源分配策略[42]，[43]。 Warneke和Leng [44]提出了一种提高内存利用率的方法。 Shojafar等。 [45]和Bari等。 [46]提出了一种动态重构网络数据中心的计算加通信资源以提高资源利用率的方法。 Reiss等。 [47]分析了Google跟踪数据，其结果对资源调度程序很有帮助。此外，研究人员还提出了一种基于随机奖励网络（SRN）的分析模型，该模型可以在不同的工作条件下设置数据中心参数[48]。

文献中有一些有关预测工作量行为的著作。这些工作预测了资源需求，资源利用率或作业/任务长度，以进行供应或调度[49]，[50]。他们提出了许多方法来估计应用程序在资源方面的未来需求，并预先分配它们，并在不需要时释放它们。但是，仅由于某些原因某些服务器可能会发生故障，并且仅提供一些解决方案是不够的，因此无法确保高效的数据中心。 Chalermarrewong等。 [51]使用ARMA（自动回归移动平均线）和故障树分析来预测在线故障。 Sedaghat等。 [52]提出了一个随机的和相关的故障存在的云数据中心的工作可靠性统计模型。 Itani等。 [53]提出了一种节点故障的解决方案。

尽管已经在传统的工作负载表征，服务器整合和数据中心中的VM迁移方面进行了研究，但很少有数据中心中共同分配的作业的工作负载表征。有一些分析阿里巴巴的跟踪数据的工作[54] \_ [59]。他们专注于云中的不平衡现象。在本文中，我们以新的维度（包括故障模式以及CPU与内存之间的相关性）对在线服务和批处理作业的共同分配来表征当代IDC。我们在本文中的发现可以帮助数据中心操作员更好地了解工作负载特征，并实现工作负载驱动的作业调度和工作负载放置。

七。结论

当代的云服务提供商巨头共同在同一集群上共同分配在线服务和批处理作业，以提高机器利用率并降低能源成本。但是，在线服务和批处理作业的混合也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。对IDC的先行操作有深刻的了解，这些IDC可以共同分配在线服务和批处理作业，可以帮助社区为IDC建立更强大的容错调度程序。

在本文中，我们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。我们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。我们在这里提供的发现和见解可以帮助数据中心操作员更好地了解工作负载特征，提高资源利用率和故障恢复能力。