(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105607967 A
(43)申请公布日 2016.05.25

(21)申请号 201510846316.7

(22)申请日 2015.11.27

(71)申请人 北京航空航天大学
地址 100191 北京市海淀区学院路37号

(72)发明人 吴文峻 蒲彦均 孙昊昊 赵德栋

(74)专利代理机构 北京科迪生专利代理有限责任公司 11251
代理人 成金玉 孟卜娟

(51)Int.Cl.
G06F 11/14(2006.01)

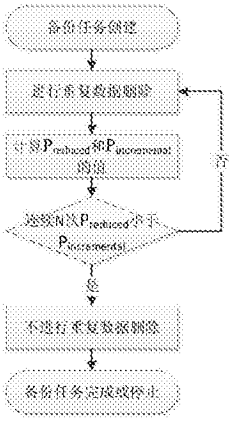
权利要求书2页 说明书4页 附图1页

(54)发明名称

一种面向数据中心基于能耗感知的数据备份方法

(57)摘要

本发明涉及一种面向数据中心基于能耗感知的数据备份方法，属于大数据节能备份技术领域：备份任务一经创建，并开始备份，全部需要经过重复数据删除过程；经过重复数据删除的备份数据集，记录本次重删的参数；计算每次重复数据删除过程后$P_{reduced}$和$P_{incremental}$的值，$P_{reduced}$表示备份过程采用重复数据删除技术而减少的耗，$P_{incremental}$表示备份过程采用重复数据删除技术而增加的能耗；并记录在数据集元数据中，并判断$P_{reduced}$和$P_{incremental}$的大小关系；如果数据集连续3次出现$P_{reduced}<P_{incremental}$的关系，则该数据集不再进行重删，即数据集按照正常流程备份；反之，只要数据集未出现连续3次存在$P_{reduced}<P_{incremental}$的情况，则继续去重并检测；无论进行重删与否，备份任务会照常执行到自然结束或被终止。本发明达到了能耗优化的目的。



CN 105607967 A

1.一种面向数据中心基于能耗感知的数据备份方法，其特征在于包括以下步骤：

(1)备份任务一经创建，即开始备份，全部备份任务均需要经过重复数据删除过程，得到重复数据删除的备份数据集；

(2)对重复数据删除的备份数据集，记录本次重删的备份数据集各项参数，所述参数包括尺寸、重删比、备份时间；其中尺寸分为原始数据大小，即逻辑尺寸和经过重复数据删除过程之后存储在介质上的数据大小，即物理尺寸；重删比是衡量重复数据删除效果最直观的数据，即备份数据集的逻辑尺寸与物理尺寸之比；以上三个参数用于在步骤(3)中计算中间变量；

(3)计算每次重复数据删除过程后$P_{reduced}$和$P_{incremental}$的值，并记录在备份数据集元数据中，并判断$P_{reduced}$和$P_{incremental}$的大小关系，其中$P_{reduced}$表示备份过程采用重复数据删除技术而减少的能耗，$P_{incremental}$表示备份过程采用重复数据删除技术而增加的能耗；

(4)如果一个数据集连续3次出现$P_{reduced}<P_{incremental}$的关系，则该备份数据集不再进行重删，即备份数据集按照正常流程备份；反之，只要数据集未出现连续3次存在$P_{reduced}<P_{incremental}$的情况，则继续去重并检测；

(5)无论进行重删与否，备份任务会照常执行到自然结束或被终止。

2.根据权利要求1所述的面向数据中心基于能耗感知的数据备份方法，其特征在于：所述步骤(1)中的重复数据删除过程的次数小于3。

3.根据权利要求1所述的面向数据中心基于能耗感知的数据备份方法，其特征在于：所述步骤(2)中$P_{incremental}$的计算公式如下：

$$P_{incremental}=\Sigma_{nodes}((P_{dedup}-P_{standby})*T_{dedup} \quad (1.1)$$

其中，T_{dedup}为重复数据删除在其上耗费的时间，$(P_{dedup}-P_{standby})$表示，在这一段时间内，由于进行重复数据删除过程，服务器功率增加的部分，P_{dedup}表示进行重复数据删除过程后存储设备的功率，$P_{standby}$是处于空闲状态下的存储设备功率。

4.根据权利要求3所述的面向数据中心基于能耗感知的数据备份方法，其特征在于：所述$(P_{dedup}-P_{standby})$或将$(P_{dedup}-P_{standby})$视为服务器功率增加的平均值，并通过能耗模型：

$$P_{active}=P_{standby}+P_{seq}+P_{rand} \quad (1.2)$$

和重复数据删除过程的资源使用率情况，计算得出，其中Pseq表示顺序IO产生的能耗，Prand表示随机IO产生的能耗，它们均为由于IO开销带来的负载相关能耗。

5.根据权利要求1所述的面向数据中心基于能耗感知的数据备份方法，其特征在于：所述步骤(1)中的$P_{reduced}$的计算公式如下：

$$P_{reduced}=(P_{MaxSeqIO}-P_{standby})*T_{iosave}*N_{active}+(P_{standby}-P_{idle})*T_{backup}*\frac{S_{dup}}{S_{energy_unit}} \quad (1.3)$$

$(P_{MaxSeqIO}-P_{standby})$是处于传输状态下与处于空闲状态下的存储设备功率之差，$P_{MaxSeqIO}$表示处于传输状态下的存储设备功率，$P_{standby}$表示处于空闲状态下的存储设备功率；T_{iosave}表示由于数据量削减而减少的传输时间，在实际实验过程中记录即可；N_{active}表示在传输过程涉及到的存储设备的数量；$(P_{standby}-P_{idle})$表示处于待机下与处于低功耗或者关闭状态下的存储设备功率之差，P_{idle}表示处于低功耗或者关闭状态下的存储设备功率；T_{backup}表示备份任务的执行周期，在备份任务下一次执行之前经过的时间，通过备份任务的设置信息得

到；$\frac{S_{dup}}{S_{energy_unit}}$表示能被空闲的设备的数量，其中$S_{energy_unit}$等同于存储系统可以操作的最小存储单元的容量，$S_{dup}$表示能被空闲的设备的容量，$S_{energy_unit}$等同于存储系统可以操作的最小存储单元的容量。

一种面向数据中心基于能耗感知的数据备份方法

技术领域

[0001] 本发明涉及一种面向能耗优化的备份方法，属于大数据节能备份技术领域。

背景技术

[0002] 在数据中心的数据备份方法中，重复数据删除的效率主要通过数据集经过重删之前和之后的占用空间之比来决定，这个比值通常被称为压缩比或者去重比。而重复数据删除的性能主要是通过数据集在备份或者存储过程中总体的读写速率进行评价的。

[0003] 目前存在的备份方法有面向云备份服务的备份方法。此方法有一个专门检查文件修改状况并做出相应处理的模块(中间件)，并能够嵌入到现有备份系统中去。并且可以根据数据集中文件的不同类别，设计了一个智能的重复数据删除系统。该系统集合了不同力度的数据划分方法，包括单一实例(SIS)，固定块划分(fixed-sized partion，FSP)，基于内容划分(CDC)和滑动块划分(sliding block)。它能够根据文件的类型，如文本文档、音频文件、图像文件、linux源码文件等，选择划分方法，以取得最好的去重比例。

[0004] 此外还有面向吞吐率的备份策略。这个策略将重删比和备份时间纳入衡量指标的范围，但是这些指标并不能直接体现出备份任务的能源消耗量。尽管一个拥有高重删比和低时间开销，即高DET值得备份数据集，在备份过程中一定也拥有较高的能耗效率。

[0005] 但是，以上的备份方法并不能准确区分这些备份任务在能耗上的表现。因此，需要制定一个从能耗角度出发，基于能耗感知的备份策略。

发明内容

[0006] 本发明的技术解决问题是：克服现有技术的不足，提供一种面向数据中心基于能耗感知的数据备份方法，以达到能耗优化的目的。

[0007] 本发明技术解决方案：一种面向数据中心基于能耗感知的数据备份方法，其特点在于包括以下步骤：

[0008] (1)备份任务一经创建，即开始备份，全部备份任务均需要经过重复数据删除过程，得到重复数据删除的备份数据集；

[0009] (2)对重复数据删除的备份数据集，记录本次重删的备份数据集各项参数，所述参数包括尺寸、重删比、备份时间；其中尺寸分为原始数据大小，即逻辑尺寸和经过重复数据删除过程之后存储在介质上的数据大小，即物理尺寸；重删比是衡量重复数据删除效果最直观的数据，即备份数据集的逻辑尺寸与物理尺寸之比；以上三个参数用于在步骤(3)中计算中间变量；

[0010] (3)计算每次重复数据删除过程后$P_{reduced}$和$P_{incremental}$的值，并记录在备份数据集元数据中，并判断$P_{reduced}$和$P_{incremental}$的大小关系，其中$P_{reduced}$表示备份过程采用重复数据删除技术而减少的能耗，$P_{incremental}$表示备份过程采用重复数据删除技术而增加的能耗；

[0011] (4)如果一个数据集连续3次出现$P_{reduced}<P_{incremental}$的关系，则该备份数据集不再进行重删，即备份数据集按照正常流程备份；反之，只要数据集未出现连续3次存在$P_{reduced}$

$<P_{incremental}$的情况，则继续去重并检测；

[0012] (5)无论进行重删与否，备份任务会照常执行到自然结束或被终止。

[0013] 所述步骤(1)中的重复数据删除过程的次数小于3。

[0014] 所述步骤(2)中$P_{incremental}$的计算公式如下：

[0015] $P_{incremental}=\Sigma_{nodes}((P_{dedup}-P_{standby})*T_{dedup}$ (1.1)

[0016] 其中，T_{dedup}为重复数据删除在其上耗费的时间，$(P_{dedup}-P_{standby})$表示，在这一段时间内，由于进行重复数据删除过程，服务器功率增加的部分，P_{dedup}表示进行重复数据删除过程后存储设备的功率，$P_{standby}$是处于空闲状态下的存储设备功率。

[0017] 所述$(P_{dedup}-P_{standby})$或将$(P_{dedup}-P_{standby})$视为服务器功率增加的平均值，并通过能耗模型

[0018] $P_{active}=P_{standby}+P_{seq}+P_{rand}$ (1.2)

[0019] 和重复数据删除过程的资源使用率情况，计算得出，其中Pseq表示顺序IO产生的能耗，Prand表示随机IO产生的能耗，它们均为由于IO开销带来的负载相关能耗。

[0020] 所述步骤(1)中的$P_{reduced}$的计算公式如下：

[0021]

$$P_{reduced}=(P_{MaxSeqIO}-P_{standby})*T_{iosave}*N_{active}+(P_{standby}-P_{idle})*T_{backup}*\frac{S_{dup}}{S_{energy_unit}} \quad (1.3)$$

[0022] $(P_{MaxSeqIO}-P_{standby})$是处于传输状态下与处于空闲状态下的存储设备功率之差，$P_{MaxSeqIO}$表示处于传输状态下的存储设备功率，$P_{standby}$表示处于空闲状态下的存储设备功率；T_{iosave}表示由于数据量削减而减少的传输时间，在实际实验过程中记录即可；N_{active}表示在传输过程涉及到的存储设备的数量；$(P_{standby}-P_{idle})$表示处于待机下与处于低功耗或者关闭状态下的存储设备功率之差，P_{idle}表示处于低功耗或者关闭状态下的存储设备功率；T_{backup}表示备份任务的执行周期，在备份任务下一次执行之前经过的时间，通过备份任务的设置信息得到；$\frac{S_{dup}}{S_{energy_unit}}$表示能被空闲的设备的数量，其中$S_{energy_unit}$等同于存储系统可以操作的最小存储单元的容量，$S_{dup}$表示能被空闲的设备的容量，$S_{energy_unit}$等同于存储系统可以操作的最小存储单元的容量。

[0023] 本发明与现有技术相比的优点在于：

[0024] (1)完全以吞吐率作为筛选的指标并不能准确区分这些备份任务在能耗上的表现。制定一个从能耗角度出发，本发明基于能耗感知的备份策略能够达到能耗优化的目的。综合两个方面的影响进行分析，找到能耗增加与能耗减少的平衡点，是制定基于能耗感知的备份策略的关键所在。根据计算得到的能耗变化值，仿照面向吞吐率的备份策略，对备份任务的数据集进行筛选。然后，备份系统只对筛选通过的数据集进行重删操作，以达到能耗优化的目的。

[0025] (2)图2展示了备份系统在不同策略影响下的平均功耗表现，由于避免了重删过程消耗的能源，同时利用空闲的硬盘进行节能，应用本发明面向能耗优化的备份方法的实验组耗能最低。相比于面向吞吐率的备份策略减少的约8％的能耗，能够关闭更多硬盘的面向能耗优化的策略降低了300W的功率。当数据集尺寸更大，实验设备规模更大时，节能效果将

更加明显。

附图说明

[0026] 图1为本发明方法实现流程图；

[0027] 图2为不同数据集在不同策略下备份系统的平均功耗。

具体实施方式

[0028] 如图1所示，为面向能耗优化的备份决策算法流程图

[0029] 由于重复数据删除技术对备份过程在能耗方面拥有两面性，因此，综合两个方面的影响进行分析，找到能耗增加与能耗减少的平衡点，是制定基于能耗感知的备份策略的关键所在。所谓能耗平衡点(energy balance point)指的是，重复数据删除过程造成备份服务器端增加的能耗开销与存储端由于空间需要减少而节约的能耗，二者相等的情况下，备份数据集的尺寸、重删比等参数状态。要找到数据集的能耗平衡点，需要计算该数据集在经过重复数据删除过程之后，在服务器端增加的能耗与在存储系统端减少的能耗。

[0030] 服务器端增加的能耗主要来自重删过程额外的时间开销和额外的计算资源开销。与一般的备份过程相比，重复数据删除在更多的时间内占用了更多的计算资源。根据服务器能耗模型，可以得到计算这一部分增加的能耗的公式：

[0031] $P_{incremental}=\Sigma_{nodes}((P_{dedup}-P_{standby})*T_{dedup}$ (1.4)

[0032] 用$P_{incremental}$表示备份过程采用重复数据删除技术而增加的能耗。一个备份任务可能需要多台服务器进行处理，需要根据情况分别计算并汇总。而对于每台服务器，重复数据删除在其上耗费的时间是T_{dedup}。这个时间可以通过数据集的重删比和服务器吞吐率由公式进行计算得到重删比即备份数据集的逻辑大小与物理大小之比。

[0033] $(P_{dedup}-P_{standby})$表示，在这一段时间内，由于进行重复数据删除过程，服务器功率增加的部分。由实验结果可以得到，使用提出的重复数据删除框架进行重删，在服务器没有其他计算密集型任务的情况下，它的资源使用情况是稳定的。因此，为了方便计算，将$(P_{dedup}-P_{standby})$视为服务器功率增加的平均值，并通过能耗模型(1.2)和重复数据删除过程的资源使用率情况，计算得出结果。

[0034] 存储系统减少的能耗可以分为两个部分，一部分是由于重删减小的数据集尺寸导致的传输开销减少带来的能耗降低，另一部分是占用空间的削减导致更多可能的存储设备可以处于低功耗运行或者关闭状态，从而节省的能耗。根据服务器能耗模型，得到计算这一部分增加的能耗的公式：

[0035]

$$P_{reduced}=(P_{MaxSeqIO}-P_{standby})*T_{iosave}*N_{active}+(P_{standby}-P_{idle})*T_{backup}*\frac{S_{dup}}{S_{energy_unit}} \quad (1.5)$$

[0036] 用$P_{reduced}$表示备份过程采用重复数据删除技术而减少的能耗。传输过程减少的开销实际上是由于将相关硬盘的传输工作时间减少了导致的。$(P_{MaxSeqIO}-P_{standby})$是处于传输状态下与处于空闲状态下的存储设备功率之差。这些存储设备需要根据具体存储系统和数据集尺寸而定。一般来讲，这些随负载有明显变化的存储设备主要是硬盘，可以通过硬盘的

能耗模型计算得到，具体公式(1.2)。

[0037] 为了简化情景，认为这里传输过程中的硬盘处于最大顺序IO负载的情况下。T_{iosave}表示由于数据量削减而减少的传输时间，可以通过数据集尺寸，去重率和存储系统带宽计算得到。N_{active}表示在传输过程涉及到的存储设备的数量，这里指涉及到的硬盘的数量。$(P_{standby}-P_{idle})$表示处于待机下与处于低功耗或者关闭状态下的存储设备功率之差。具体采用哪个状态进行计算，要根据存储系统能够控制这些设备的哪些状态转换而决定。T_{backup}表示备份任务的执行周期，在备份任务下一次执行之前经过的时间，可以通过备份任务的设置信息得到。$\frac{S_{dup}}{S_{energy_unit}}$表示能被空闲的设备的数量，其中$S_{energy_unit}$等同于存储系统可以操作的最小存储单元的容量。根据存储系统的设置和功能不容，它能控制的设备最小单元可能是一块硬盘、多块硬盘形成的RAID组或者是包含相应控制服务器的存储子系统。

[0038] 通过上面的公式(1.6)可以计算得到由于在备份过程中引入了重复数据删除技术而产生的能耗的改变。从能耗角度出发，可以通过这个能耗变化预测在实际备份过程中对备份任务数据集进行重复数据删除，而产生的对包括存储系统和备份服务器在内的整个备份系统的影响。根据这个计算得到的能耗变化值，仿照面向吞吐率的备份策略，对备份任务的数据集进行筛选。然后，备份系统只对筛选通过的数据集进行重删操作，以达到能耗优化的目的。这里仍然沿用之前的备份轮次阈值N，在本发明提出的备份策略中，供预测重删比的备份轮次N指定为3。这是因为能耗预测同样需要对数据集的重删比进行预测，需要经过一定次数的重删过程，对数据集特征参数进行统计。设定了阈值P，并根据阈值P与差值$(P_{reduced}-P_{incremental})$的大小关系。

[0039] 为了使经过筛选的数据集在执行重复数据删除过程时有更高的能耗效率，在具体实现中，将阈值P设置为0。这样做的目的是为了保证经过重复数据删除的数据集一定会因此降低备份系统的能源开销。类似面向吞吐率的备份策略，论文将面向能耗优化的备份策略的流程定义如下，如图1所示：

[0040] (1)备份任务一经创建，并开始备份，全部需要经过重复数据删除过程。

[0041] (2)经过重复数据删除的备份数据集，记录本次重删的各项参数，包括尺寸、重删比、备份时间等。并根据前文重删比估计方法，计算重删比，即备份数据集的逻辑大小与物理大小之比。

[0042] (3)计算每次重复数据删除过程后$P_{reduced}$和$P_{incremental}$的值，并记录在数据集元数据中。并判断$P_{reduced}$和$P_{incremental}$的大小关系。

[0043] (4)如果一个数据集连续3次出现$P_{reduced}<P_{incremental}$的关系，则该数据集不再进行重删，即数据集按照正常流程备份。反之，只要数据集未出现连续3次存在$P_{reduced}<P_{incremental}$的情况，则继续去重并检测。

[0044] (5)无论进行重删与否，备份任务会照常执行到自然结束或被终止。

[0045] 提供以上实施例仅仅是为了描述本发明的目的，而并非要限制本发明的范围。本发明的范围由所附权利要求限定。不脱离本发明的精神和原理而做出的各种等同替换和修改，均应涵盖在本发明的范围之内。

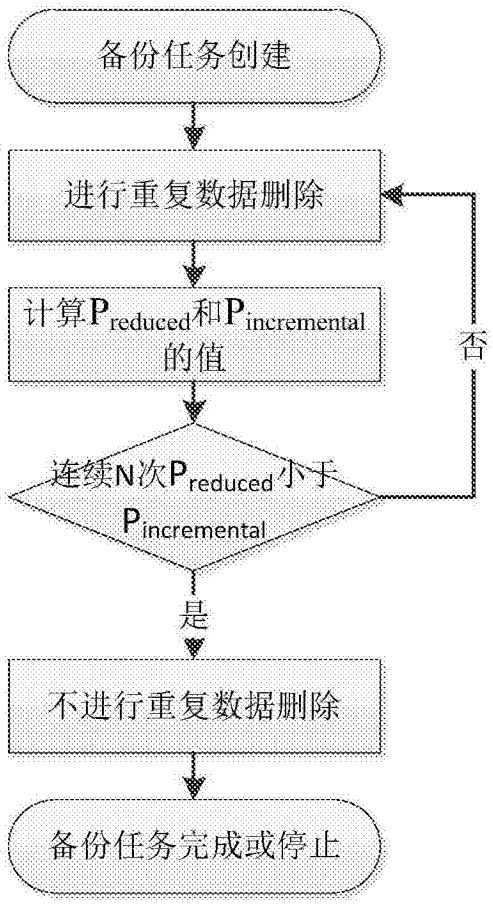

备份任务创建
进行重复数据删除
计算$P_{reduced}$和$P_{incremental}$的值
连续N次$P_{reduced}$小于$P_{incremental}$
否
是
不进行重复数据删除
备份任务完成或停止

图1

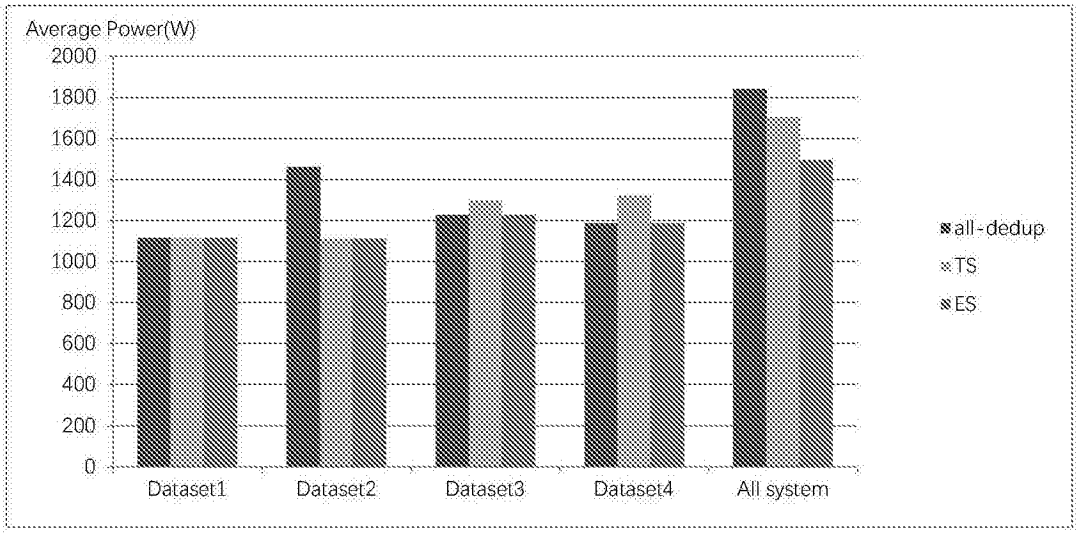

Average Power(W)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Dataset1
Dataset2
Dataset3
Dataset4
All system
all-dedup
TS
ES

图2