发明名称
用于存储控制器的高效生命周期管理的方法和系统

摘要
本发明涉及用于存储控制器的高效生命周期管理的方法和系统。用于计算机网络存储系统中的旧存储控制器的高效迁移过程的计算方法。该方法提供了组合迁移的非 pNFS 数据存储与新的临时并行 NFS 数据存储。在实施例中，该方法包括一系列消耗相对短时间的操作，其中存储系统将存储数据从存储在 pNFS 存储器存储的遗留数据的旧存储控制器迁移。其中高效的迁移实现了回收布局（pNFS，独立式 pNFS MDS）以及将存储数据重新引导至新存储控制器的能力。在另一实施例中，该方法包括操作序列，在所述操作序列下，存储系统将数据从具有非 pNFS 数据存储的存储控制器有效地迁移。在该实施例中，在引退期间的存储使用率组合了遗留的非 pNFS 存储以及新的临时 pNFS 存储空间管理。
1. 一种用于管理存储并行存取网络系统的多个第一存储设备中的至少一个中的数据对象和布局数据的计算机化方法，所述并行存取网络系统具有元数据服务器，所述元数据服务器管理所述布局数据以及将所述数据对象传递到在所述并行存取网络系统下操作的多个第二存储设备中的至少一个，包括用于最优存储容量管理以及在一段时间段中多个第一存储设备中的所述至少一个的使用的步骤序列，所述时间段与所述数据对象从多个第一存储设备中的所述至少一个传递到多个所述第二存储设备中的所述至少一个相关联，其中与多个第一组存储设备中的所述至少一个相关联的所述数据不在所述元数据服务器下管理，所述方法包括以下步骤：

定义选自一组选项的所述多个第一存储设备中的所述至少一个的期望存储容量使用率参数目标，其包括通过系统存储管理员来定义所述参数以及通过系统缺省选项来定义所述参数；

将与待租借或租赁的所述多个第一存储设备中的所述至少一个有关的新的布局数据组指定给所述系统元数据服务器；

通过测量代表所述第一存储设备组的容量使用率的周期使用率来重新计算所述多个第一存储设备中的所述至少一个的周期使用率存储容量；

计算待指定给由所述元数据服务器管理的布局池的周期自由空间参数，所述自由空间＝所述存储期望存储使用率－所述存储周期使用率；

在调整所述布局组的大小的同时，将所述存储计算周期自由空间添加到指定大小的所述布局组；

重复重新计算所述第一存储设备组的周期使用率存储容量的序列，以及

当所述系统管理员检测到在所述多个第一存储设备中的所述至少一个上仅留有不在与所述多个第一存储设备中的所述至少一个相关联的所述元数据服务器下管理的非显著量的所述对象数据和关联布局时，结束重新计算过程。

2. 根据权利要求1所述的计算机化方法，还包括以下步骤：

在重新计算所述多个第一存储设备中的所述至少一个的周期使用率存储容量之前，等待周期监视器。

3. 根据权利要求1所述的计算机化方法，还包括以下步骤：

在所述步骤序列结束时，对于所述多个第一存储设备中的所述至少一个执行引退程序。

4. 根据权利要求3所述的计算机化方法，其中所述引退程序包括以下步骤：

从其新的分配选项提取与所述多个第一存储设备中的所述至少一个相关联的布局以避免其由多个所述系统客户端中的任一个进一步用于所述系统新应用；

对于与所述多个第一存储设备中的所述至少一个相关联的任意选定布局组，阻挡新的布局请求；

向在所述选定存取数据组中的共享相关布局副本的多个客户端发出布局调用请求；

等待直到预定租赁时间以从所述客户端得到关于共享匹配布局的布局返回确认通知；

从所述多个客户端接收布局返回确认响应；

将与所述选定布局组相关联的对象数据从所述多个第一存储设备迁移到新选定的多
个存储设备；以及
重复从所述多个第一存储设备中的所述一个到所述多个第二存储设备中的所述至少一个的对象数据传递步骤序列，直到所述多个所述第二存储设备中的所述至少一个的所述内容内容到所述多个所述第一存储设备中的至少一个。
5. 根据权利要求 1 所述的计算机化方法，其中所述多个所述第一存储设备中的所述至少一个和所述多个所述第二存储设备中的所述至少一个的所述数据内容通过所述具有 MDS 数据服务器的 pNFS 网络系统。
6. 根据权利要求 5 所述的计算机化方法，其中所述多个第一存储设备中的所述至少一个包括 NAS 文件级类型的存储数据服务器。
7. 根据权利要求 5 所述的计算机化方法，其中所述多个第一存储设备中的所述至少一个包括 SAN 块级类型的存储数据服务器。
8. 根据权利要求 4 所述的计算机化方法，其中所述具有元数据服务器的所述并行存取网络系统为具有 MDS 数据服务器的 pNFS 网络系统。
9. 根据权利要求 8 所述的计算机化方法，其中所述多个所述第一和所述第二存储设备包括 NAS 文件级类型的存储数据服务器。
10. 根据权利要求 8 所述的计算机化方法，其中所述多个所述第一和所述第二存储设备包括 SAN 块级类型的存储数据服务器。
11. 一种并行存取网络文件系统，包括：
元数据服务器，其存储并管理所述数据；
多个客户端，其共享所述系统；
存储数据对象和所述的多个第一存储设备中的所述至少一个；
多个第二存储设备中的所述至少一个；以及
其中，所述系统在预期用于最优存储容量管理以及在与服务器相关联的时间段内多个第一存储设备中的所述至少一个的使用的步骤序列下执行多个第一存储设备中的所述至少一个的所述存储程序，其中所述数据对象逐渐地从所述多个第一存储设备传递到所述多个第二存储设备组，并且所述存储在所述多个第一存储设备中的所述数据不在所述元数据服务器下管理。
12. 根据权利要求 11 所述的系统，其中所述在所述多个第一存储设备中的所述所述布局在所述程序期间租借或租赁给存储并管理所述数据的所述元数据服务器。
13. 根据权利要求 12 所述的系统，其中所述的所述最优存储容量管理和所述多个第一存储设备的使用，所述元数据服务器正在使用所述租赁的布局来在所述多个第一存储设备中临时存储附加租赁数据对象。
14. 根据权利要求 13 所述的系统，其中所述元数据服务器存储所述租赁数据对象以使所述多个第一存储设备上的逐渐减小数量的所述最优存储的数据对象与所述临时租赁的数据对象之和实际上保持恒定，同时将所述多个第一存储设备的数据存储容量保持到由包括系统管理员和系统用参数的组中的一个限定的最优存储水平。
15. 根据权利要求 11 所述的系统，其中所述并行存取网络文件系统为具有 MDS 数据服务器的 pNFS 网络系统。
16. 根据权利要求 11 所述的系统，其中所述多个第一存储设备为 NAS 服务器，并且所述存储的数据对象和布局为文件和卷。
17. 根据权利要求11所述的系统，其中所述多个第一存储设备为NAS服务器，并且所述存储的数据对象和布局为块和LUN。

18. 一种用于在并行存取网络文件系统中执行多个存储设备引退程序的设备，所述并行存取网络文件系统包括存储并管理布局数据的元数据服务器，共享所述系统的多个客户端、存储数据对象和布局的多个第一存储设备中的至少一个以及多个第二存储设备中的至少一个，其中存储数据对象和布局的所述第一多个存储设备的所述引退程序是在预期用于所述第一多个存储设备的最优存储容量管理以及在与所述引退程序相关联的时间段内的使用的步骤序列下执行的，其中所述数据对象从所述第一多个存储设备传递到所述第二多个存储设备，并且其中存储在所述第一多个存储设备中的所述数据不在所述元数据服务器下管理，所述设备包括：

第一装置，其通过系统存储管理员来定义所述第一多个存储设备的期望数据存储容量使用率参数目标；

第二装置，其将与待租借或出租的所述第一多个存储设备有关的新的布局数据组指定给所述系统元数据服务器；

第三装置，其在重新计算所述第一多个存储设备的周期使用率存储容量之前等待周期监视器；

第四装置，其通过第五装置来重新计算所述第一存储设备组的周期使用率存储容量以测量代表多个所述第一存储设备的容量使用率的Periodic_utilization；

第六装置，其计算待指定给由所述元数据服务器管理的布局池的Periodic_free_space，其中Periodic_free_space=Desired_utilization-Periodic_utilization；

第七装置，其通过调整大小将所述计算的Periodic_free_space添加到指定大小的所述布局组中；

第八装置，其重复重新计算所述第一多个存储设备的周期使用率存储容量的序列，以及

第九装置，当所述系统管理员检测到在所述第一多个存储设备上仅留有不在与第一多个存储设备相关联的所述元数据服务器下管理的非显著量的所述对象数据和关联布局时，所述第九程序指令执行所述第一多个存储设备的引退程序。

19. 根据权利要求18所述的设备，其用于在所述第一多个存储设备中的至少一个上执行引退程序，进一步包括执行所述第一多个存储设备中的所述至少一个的引退程序的第十装置。
用于存储控制器的高效生命周期管理的方法和系统

技术领域

在本发明的一些实施例中，本发明涉及计算机存储数据存取和管理高级解决方案，更特别地而非排他地涉及在实现基于带外 pNFS 协议的解决方案的同时用于高效存储控制器生命周期管理的方法和系统，其中组织中遗留的文件管理系统用作数据服务，其能够在单个数据服务器上混合前 pNFS 数据和后 pNFS 数据，以提高需要收回和更换的数据服务器的停机期间使用效率。

背景技术

高性能数据中心已主动朝向类似于群集计算和多核处理器的并行技术移动。尽管这种提高的并行性能使用克服了大部分的计算瓶颈，但是其将性能瓶颈转移到存储 I/O 系统。为了确保计算群集实现最大性能，存储系统必须针对并行性进行优化。当与大规模、高性能计算群集结合实现时，工业标准网络联接存储 (NAS) 架构具有严重的性能瓶颈和管理挑战性。通过允许计算客户端对存储设备进行直接读和写，完全消除了文件管理程序头瓶颈以及利用专用协议来允许单个文件系统容量和性能线性地扩大至极高的水平，并行存储采取了极不同的方法。

近年来，客户端的存储输入和 / 或输出 (I/O) 带宽要求已经快速地超过了网络文件服务器向量其供给的能力。在根据网络文件系统 (NFS) 协议运行的设备中遭遇该问题。传统的 NFS 架构包括置于磁盘驱动器前而且由 NFS 导出文件系统的文件管理程序头。在典型的 NFS 架构下，当客户端试图存取文件时，当大量客户端想要同时存取数据或者如果数据集增长过大，则情况变得复杂。然后，NFS 服务器迅速变成瓶颈并且严重影响系统性能，因为 NFS 服务器定位在客户端计算机和物理存储设备之间的数据路径中。

为了克服该问题，已经开发了并行 NFS (pNFS) 协议和相关的系统存储管理架构。pNFS 协议及其支撑架构允许客户端直接地并行地对存储设备进行存取。与之前的 NFS 架构相比，pNFS 架构提高了可伸缩性和性能。该提高是通过数据和元数据的分离以及利用数据路径外的元数据控制器来实现的。

在使用时，pNFS 客户端发出对元数据服务请求的数据控制请求，并且随后且同时地调用对数据服务器群集的多个数据存取请求。不同于通过单个 NFS 存储服务器处理数据控制请求和数据存取请求的常规 NFS 环境，pNFS 配置支持所需的多个数据服务器以服务客户端请求。因此，pNFS 配置能够用于大幅提高常规 NFS 存储系统的可伸缩性。pNFS 的协议规范可在 URL: www.ietf.org 找到，参见 NFS4.1 标准，在 URL: www. open-nNFS. org 和 www. ietf.org, Requests for Comments (RFC) 5661-5664，其包括从基础协议和协议扩展版保留的特征。RFC 5661-5664 包括主要扩展，诸如会话、目录代表外部数据表示标准 (XDR) 描述，与 NFSv4.1 协议一起使用的基于块的布局型定义的规格以及与 NFSv4.1 协议一起使用的基于对象的布局型定义。

在将计算机存储系统升级至 pNFS 环境的同时，尤其而非唯一重要地使共享 NFS 存储控制器引退在许多生产 / 操作环境中花费数月。关断控制器需要存储数据的迁移以及相
应地更新所有客户端的应用。该过程花费相当长的时间，原因如下：
[0007] 1. 尽管控制器知其所保持的数据，但它们不知道当前使用该数据的客户端
应用，或者可能在额外的时间使用该数据。
[0008] 2. 在管理员知道使用应用的情况下，同步以及就其所使用的停工时间
隙达成一致意见都花费时间。
[0009] 上述存储控制器的长的停工时间要求过程对于存储域网络（SAN）以及网络化附接
存储（NAS）控制器（也称为阵列（SAN）或文件管理程序（NAS））两者而言是真实的。
[0010] 存在多种方法来克服基本上较长的控制器的停工时间过程局限。一个这样的示例
性的已知解决方案基于以下方法：
[0011] 一旦管理员识别出相关应用及其数据，实现以下步骤；
[0012] a. 为应用调度停工时间窗；
[0013] b. 数据从旧的将要引退的控制器复制到新的控制器。在旧的和新的控制器支
持相同的专用同步镜像协议的具体方案中，所述复制能够在停工时间之前完成；以及
[0014] c. 应用转下（bring down），其存储被重构，然后应用重新启动。也就是说，在高
级虚拟基础架构上运行的应用可利用不同的存储迁移到另一群集，同时保留系统操作持续
性。
[0015] 利用将要引退的控制器对于所有识别的已用的每个重复该过程。当管理员给定
它们完成后，管理员通常监控将要引退控制器上的 I/O 数据通信量以查看是否存在活跃的
请求。如果一段时间暂时无活动可见，则控制器被视为空闲。
[0016] 现有的停工时间过程解决方案的一些已知缺陷可概括如下：a. 对于应用，停工时
间花费大量的时间；以及 b. 永不存在所有客户端应用都知道数据状态变化的完全程
度的确定性。因此，旧有的控制器保持活跃数月，从而识别尽可能多的客户端应用。同时，
存储控制器消耗资源并且以极低的使用率运行。图 1 示例了示例性的未用尽的控制器，其
在 1 月份启动引退过程并且保持活跃 9 个月直到最终关断。
[0017] 因此，本领域对于 pNFS 存储系统在与老旧控制器引退过程有关的引退周期的
持续时间的情况或者可选地对于这样的非 pNFS 存储系统的情况存在需求，以在相当长的
低使用率时间段内提高将要引退的存储控制器的使用率，直到其断开，同时以其全部能
力持续地操作和管理系统操作数据处理吞吐量和性能。
[0018] 词汇表
[0019] 网络文件系统（NFS）- 分布式文件系统开放标准协议，其允许客户端计算机上的
用户通过网络访问文件，方式与客户端计算机上的用户如何访问本地存储的方式相似。
[0020] NFSv4-NFS 版本 4，包括性能提高和更强的安全性。其支持群集服务器部署，包括
提供对分布在多个服务器之中的文件的可伸缩并行访问的能力（NFS 扩展）。
[0021] 并行 NFS（pNFS）- NFS v4.1 的部分，其允许计算客户端直接且并行地访问存储设
备。pNFS 架构通过数据和元数据的分离以及将元数据服务器移出数据路径来消除与 NFS 服
务器群集的可伸缩性和性能问题。
[0022] pNFS 元数据服务器（MDS）- 专用服务器，其在 pNFS 协议下发起并管理对数据服
务器群集的数据控制和访问请求。
[0023] 网络文件服务器 - 附接至网络的计算机工具，其具有提供共享磁盘访问的位置的
主要用途，即，能够由附接至同一计算机网络的工作站访问的计算机文件的共享存储。文件服务器不意在执行计算任务，且不表示其客户端运行程序。文件服务器被设计为主要使用在通过工作站实施计算的同时进行数据存储和检索。

[0024] 外部数据表示(XDR) - 标准数据序列化格式，用于诸如计算机网络协议。其允许数据在不同类型的计算机系统之间传递。从本地转换成 XDR 被称为编码。从 XDR 转换成本地表示被称为解码。XDR 实现为软件功能库，其在不同的操作系统之间便携且独立于传输层。

[0025] 存储域网络(SAN)（也称为阵列）- 提供对合并的、块级计算机数据存储的存取的专用网络。SAN 主要用于使诸如磁盘阵列的存储设备对于服务器可存取以使设备看起来好像是与操作系统本地附接的设备。SAN 通常具有其自身的存储设备网络，其通常不能由其它设备通过局域网存取。SAN 不提供文件抽象(file abstraction)，仅提供块级操作。构建在 SAN 之上的提供文件级存取的文件系统称为 SAN 文件系统或共享磁盘文件系统。

[0026] 网络附接存储(NAS)（也称为文件管理程序）- 文件级计算机数据存储，其与提供对客户端的异类组的数据存取的计算机网络连接。NAS 作为文件服务器而运行，通过其硬件、软件或那些元件的构造而专用于该任务。NAS 网络提供作为计算机工具。用于存储和维护文件的专用计算机。NAS 是在多个计算机之间共享文件的便利方法。与文件服务器相比，其有益于网络附接存储，包括更快的存取、更容易的管理以及简单的配置。

[0027] NAS 系统 - 网络工具，其包含一个或多个磁盘驱动器，通常设置成逻辑的、冗余的存储容或 RAID。网络附接存储免除了文件从网络上的其它服务器服务的负担。它们通常利用诸如 NFS、SMB/CIFS 或 AFP 的网络文件共享协议来提供对文件的存取。

[0028] 独立磁盘的冗余阵列(RAID) - 将多个磁盘驱动部件组合到逻辑单元中的存储技术。根据所要求的冗余级别和性能，数据以多种方式中的称为“RAID 级别”的一种方式分布在驱动器上。RAID 用作用于计算机数据存储方案的术项，其能够在多个物理驱动器之间划分和复制数据。RAID 为存储虚拟化的示例，并且阵列能够作为一个单—驱动器由操作系统来存取。

[0029] 逻辑单元号(LUN) - LUN 能够用于将磁盘存储的较大或较小的视图提供给服务器。在 SAN 存储环境中，LUN 表示逻辑抽象，或者物理磁盘设备 / 存储卷和应用程序之间的虚拟层。用于存储器的基本存储单元称为 LUN。每个 LUN 标识具体的逻辑单元，其可以为硬盘驱动器的一部分、整个硬盘或存储设备中的多个硬盘。LUN 可指代整个 RAID 组、单个磁盘或分区、或多个硬盘或分区。逻辑单元就像作为一个单个设备一样被处理。

[0030] 逻辑卷卷 - 逻辑卷由一个或多个逻辑驱动器构成，成员逻辑驱动器能够为相同的 RAID 级或不同的 RAID 级。逻辑驱动器仅为独立物理驱动阵列。逻辑驱动器对于主机而言与本地硬盘驱动器相同。逻辑卷能够划分成最大 8 个分区。在操作期间，主机将非分区的逻辑卷或分区的逻辑卷的分区视为一个单一物理驱动器。

[0031] 客户端 - 负于网络上的多用户计算机或终端的术语。客户端登录到服务器上的网络中并且容许使用网络上的资源。客户端计算机通常较慢且要求网络的容许，这将它们与服务器计算机区分开。

[0032] 布局 - 指定给应用程序或在存储系统存储器中包含具体数据包的位置的客户端的存储区域。
发明内容
[0033] 下面的实施例及其方案结合方法和系统进行描述和说明，其旨在示例性的和说明性的，而非限制范围。在各个实施例中，所述问题中的一个或多个已经减轻或消除，而其它实施例涉及其它优点或改进。
[0034] 因此，在本领域普通认为在使共享NFS存储控制器引入过程中存在需求，在pNFS环境下操作的本发明实施例中的一个中，便能实质上缩短将要引入的pNFS存储控制器的引入时间，直到其能够停止运行。同时仍以其全部性能操作和管理系统数据管理操作吞吐量。
[0035] 在pNFS环境下操作的本发明方法的一个实施例中，克服了现有技术中将要引入的存储控制器的长时间段的低使用率的局局性。这是通过如下实现的：对虚拟化进行杠杆作用调节以及实现其他网络文件系统(NFS)协议的pNFS版本以实质上缩短整个控制器引入期间周期所需的时间，从而避免现有技术中在停工时间周期内将要引入的存储控制器的使用期间下限的持续时间。停工时间周期的显著缩短是通过依赖于两个pNFS环境相关的副产物而得以支持的。a. pNFS固有的数据和元数据的分离以及利用数据路径之外的元数据服务器(MDS)；以及b. 大多数pNFS布局类型(例如，块、NFS对象、伸缩文件(flex-file))具有使用遗留的文件管理程序或阵列作为其数据服务器(DS)的能力。
[0036] 因此，在本领域普通认为在使共享NFS存储控制器引入过程中存在需求，在非pNFS环境下尤其重要的是同时升级至pNFS环境的同时，或在混合的非pNFS和pNFS系统环境下操作的另一实施例中，使能NFS存储控制器的组织引退的引退阶段内提高将要引退的存储控制器的最优使用，直到其能够停止运行。因此，本发明另一实施例将支持更好的维护以及系统数据管理操作吞吐量的优化操作和管理至其全部性能。
[0037] 本发明方法的第二实施例克服了现有技术中在非pNFS系统环境下将要引退的存储控制器的低使用率的局局性。这是通过对虚拟化进行杠杆作用调节以及通过实现普通网络文件系统(NFS)协议的pNFS版本以避免在停工时间周期内将要引退存储控制器的低使用率来实现的，有赖于两个pNFS环境相关的副产物：a. pNFS固有数据和元数据的分离以及使用数据路径之外的元数据服务器(MDS)；以及b. 大多数pNFS布局类型(例如，块、NFS对象、伸缩文件)具有使用遗留的文件管理程序或阵列作为其数据服务器(DS)的能力。
[0038] 因此，提供了一种用于管理存储在并行存储网络系统中的多个第一存储设备中的至少一个中的数据对象和布局数据的计算方法，所述并行存储网络系统具有元数据服务器，所述元数据服务器管理所述布局数据以及将所述数据对象传递到在所述并行存储网络系统下操作的多个第二存储设备中的至少一个，包括用于最优存储容量(capacity,性能)管理以及在与所述数据对象从多个第一存储设备中的所述至少一个传送到多个所述第二存储设备中的所述至少一个的使用的步骤序列，其中与多个第一存储设备中的至少一个相关联的所述数据不在所述元数据服务器下管理。所述方法包括以下步骤：
[0039] 定义选择一组选项的所述多个第一存储设备中的至少一个的期望存储容量使用率参数目标，包括通过系统存储管理员来定义所述参数以及通过系统缺省选项来定义所述参数；
将与待租借或租赁的所述多个第一存储设备中的所述至少一个有关的新的布局数据组指定给所述系统元数据服务器；
通过测量代表所述第一存储设备组的容量使用率的周期使用率来刷新计算所述多个第一存储设备中的所述至少一个的周期使用率存储容量；
计算待指定给由所述元数据服务器管理的布局池的周期自由空间参数，其中所述存储周期自由空间 = 所述存储期望存储使用率 - 所述存储周期使用率；
在调整所述布局组的大小的同时，将所述存储计算周期自由空间添加到指定大小的所述布局组；
重复刷新计算所述第一存储设备组的周期使用率存储容量的序列；以及
当所述系统管理员检测到在所述多个第一存储设备中的所述至少一个上仅留有不显著量的所述对象数据，所述多个第一存储设备中的所述至少一个相关联的所述元数据服务器下管理的非显著量的所述对象数据和相关布局时，结束刷新计算过程。
此外，该方法还包括以下步骤：在重新计算所述多个第一存储设备中的所述至少一个的周期使用率存储容量之前，等待周期监视器。
此外，该方法还包括以下步骤：在所述步骤序列结束后时，对于所述多个第一存储设备中的所述至少一个执行引退程序。
此外，引退程序包括以下步骤：
从其新的指定选项提取与所述多个第一存储设备中的所述至少一个相关联的布局以避免其由多个所述系统客户端中的任一个进一步用于所述系统的应用。
对于与所述多个第一存储设备中的所述至少一个相关联的任意选定布局组，阻断新的布局请求；
向在所述选定存取数据组中的共享相关布局副本的多个客户端发出布局调用请求；
等待达到预定租借时间以便所述客户端得到关于共享匹配布局的布局返回确认通知；
从所述多个客户端接收布局返回确认响应；
将与所述选定布局组相关联的对象数据从所述多个第一存储设备迁移到新选定的多个存储设备；以及
重复从所述多个第一存储设备中的所述至少一个到所述多个第二存储设备中的所述至少一个的对象数据传递步骤序列，直到所述多个所述第一存储设备中的所述至少一个的全部数据内容传递到所述多个所述第二存储设备中的所述至少一个。
此外，具有元数据服务器的所述并行存取网络系统为具有 MDS 数据服务器的 pNFS 网络系统。
此外，第一存储设备和第二存储设备可以包括 NAS 文件级类型的存储数据服务器或 SAN 块级类型的存储数据服务器。
此外，具有元数据服务器的并行存取网络系统为具有 MDS 数据服务器的 pNFS 网络系统。
另外，提供一种并行存取网络文件系统，包括：元数据服务器，其存储并管理布局数据；多个客户端，其共享所述系统；多个第一存储设备中的至少一个，其存储数据对象和
布局；多个第二存储设备的至少一个；以及其中，所述系统在预期用于最优存储容量管理以及在与引退程序相关联的时间段内多个第一存储设备中的所述至少一个的使用的步骤序列下执行多个第一存储设备中的所述至少一个的所述引退程序，其中所述数据对象逐渐地从所述多个第一存储设备传递到所述多个第二存储设备组，并且其中存储在所述多个第一存储设备中的所述数据对象不在所述元数据服务器下管理。

[0060] 此外，存储在所述多个第一存储设备中的所述布局在所述程序期间租借或租赁给存储并管理布局数据的所述元数据服务器。执行所述最优存储容量管理和所述多个第一存储设备的使用，所述元数据服务器正在使用所述租赁的布局来临时存储在所述多个第一存储设备的附加租赁数据对象中。

[0061] 此外，所述元数据服务器存储所述租赁数据对象以使所述多个第一存储设备上的所述最初存储的数据对象与所述临时租赁的数据对象的逐渐减小数量之和实际上保持恒定，同时将所述多个第一存储设备的数据存储容量保持到由包括系统管理员和系统缺省参数的组中的一个限定的其最优存储水平。

[0062] 此外，第一存储设备可以为 NAS 服务器，并且所存储的数据对象和布局可以为块和 LUN。

[0063] 另外，提供一种用于在并行存取网络文件系统中执行多个存储设备引退程序的引退程序的计算机程序产品，所述并行存取网络文件系统包括存储并管理布局数据的元数据服务器、共享所述系统的多个客户端、存储数据对象和布局的多个第一存储设备中的至少一个以及多个第二存储设备中的至少一个，其中存储数据对象和布局的所述第一多个存储设备的所述引退程序是在预期用于所述第一多个存储设备的最优存储容量管理以及在与所述引退程序相关联的时间段内的使用的步骤序列下执行的，其中所述数据对象从所述第一多个存储设备传递到所述第二多个存储设备，并且其中存储在所述第一多个存储设备中的所述数据不在所述元数据服务器下管理。

[0064] 所述计算机程序包括：第一程序指令，其通过系统存储管理员来定义所述第一多个存储设备的期望数据存储容量使用率参数目标；第二程序指令，其将与待租借或租赁的所述第一多个存储设备有关的新的布局数据组指定给所述系统元数据服务器；第三程序指令，其在重新计算所述第一多个存储设备的期望数据存储容量使用率存储容量之前待存储器；第四程序指令，其通过第五程序指令来重新计算所述第一存储设备的期望数据存储容量使用率存储容量以测量代表多个所述第一存储设备的期望数据存储容量使用率存储容量；第六程序指令，其计算待指定给由所述元数据服务器管理的布局池的期望自由空间，其中 Periodic_free_space=Desired_utilization−Periodic_utilization；第七程序指令，其通过调整大小将所述计算的 Periodic_free_space 添加到指定大小的所述布局组中；第八程序指令，其重复重新计算所述第一多个存储设备的期望数据存储容量的序列，以及第九程序指令，当所述系统管理员检测到在所述第一多个存储设备上仅留有在与第一多个存储设备相关联的所述元数据服务器下存储的非显著量的所述对象数据和关联布局时，所述第九程序指令执行所述第一多个存储设备的引退程序。所述第一、第二、第三、第四、第五、第六、第七和第八程序指令存储在所述计算机可读存储介质上。

[0065] 此外，提供一种用于在所述第一多个存储设备中的至少一个上执行引退程序的计算机程序产品，其中程序进一步包括执行所述第一多个存储设备中的所述至少一个的引退
程序的第十程序指令。

[0066] 除非进行定义，否则本文使用的所有技术和／或科学术语具有与本发明所属领域的普通技术人员所通常理解的含义相同的含义。虽然与本文所描述的相似或等同的方法和系统能够用于本发明的实施例的实践或测试，文模式的是示例性的方法和／或系统。在冲突的情况下，以包括定义的专利说明书为准。另外，本文中的材料，方法，系统和示例仅为示例性的，而不旨在必要地限制。

附图说明

[0067] 本文参照附图仅通过举例的方式描述了本发明的一些实施例。现在具体详细地参照附图，应强调的是，所述出的详细说明是通过举例的方式以及为了示例性的讨论本发明实施例的目的。在这方面，结合附图进行的描述使得如何可实施本发明的实施例对于本领域技术人员而言是显然的。

[0068] 图1是示出在系统管理员实施引退过程中具有未用尽数据的控制器的示例性的现有技术遗留的非pNFS，非虚拟化的存储系统的控制器使用率百分比对持续时间的示例使用率曲线图的图解。

[0069] 图2是根据本发明的一些实施例的包括元数据服务器（MDS）和在pNFS系统环境下也称为数据服务器的多个存储设备的存储系统的示意性图示。

[0070] 图3A-3E为根据本发明的一个实施例的状态机的示意性流程图图示，其中状态反射动作并且转移箭头涉及内部或外部触发，内部或外部触发是针对特定布局而执行的，其中在该状态机中展示了通过回收布局（pNFS，独立式pNFS MDS）以及将老旧数据（虚拟化存储）重新引导至新的控制器的能力完成的，仅在pNFS存储下迁移遗留数据。

[0071] 图4为本发明的实施例的在pNFS上的遗留数据加上虚拟化存储的情况下示例性的存储控制器的示例使用率曲线图的图示，其中在系统管理员实施引退过程中从未用尽数据的控制器迁移遗留数据是由于回收布局（pNFS，独立式pNFS MDS）以及将老旧数据（虚拟化存储）重新引导至新控制器的能力而在在得多的时间段内仅在pNFS存储下完成的。

[0072] 图5A-5B是根据本发明的另一实施例的状态机的示意性流程图图示，其中从具有未在pNFS存储下运行的数据的存储控制器迁移数据可视为更难，复杂且耗费时间。在该实施例中，在引退期间内的存储器使用率组合了遗留的pNFS，非虚拟化存储以及新的临时性pNFS存储空间使用。在该实施例中，我们不缩短控制器淡出和引退的时间段，而是集中于通过系统管理员使老旧控制器引退的过程所需的时间段内提高老旧数据控制器的使用率。

[0073] 图6为本发明方法的另一示例性实施例的示例的使用率曲线图的图示，其中在通过系统管理员实施引退的过程中从未用尽数据的控制器迁移遗留数据时实现该方法，并且其中，控制器在引退过程中组合遗留的pNFS存储空间数据内容以及新的临时性pNFS存储空间。这种情况被认为更加负载和耗时。在该情况下，我们不缩短控制器淡出的时间段，而是集中于通过在其上存储更多的临时pNFS加上虚拟化数据内容来提高在停机时间周期内老旧控制器使用率。
具体实施方式
[0074] 在本发明的一些实施例中，本发明涉及存取数据，更特别地而非排他地，涉及带外存取数据管理和老旧数据存储控制引退（retirement）的方法和系统。
[0075] 在详细地说明本发明的至少一个实施例之前，应理解的是，本发明的应用不必局限于下面的说明书中所阐述的和/或附图中图示的和/或实施例中的部件和/或方法的构造和布局的细节。本发明能够具有其它的实施例或者以各种方式进行实施或实施。
[0076] 本领域技术人员将理解的是，本发明的方案可具体实现为系统，方法或计算机程序产品。因此，本发明的方案可呈现为完全硬件实施例，完全软件实施例（包括固定、驻留软件、微代码等）或组合了软件和硬件方案的实施例的形式，这些在本发明中通常可称为“电路”、“模块”或“系统”。此外，本发明的方案可呈现为计算机程序产品（例如，该计算机程序产品实现为具有计算机可读程序代码的一个或多个计算机可读介质）。
[0077] 可以使用一个或多个计算机可读介质的任意组合。计算机可读介质可以为计算机可读信号介质或计算机可读存储介质。计算机可读存储介质可以为例如但不限于电子的、磁的、光的、电磁的、红外的、或半导体的系统、装置或设备或前述任意适当的组合。计算机可读存储介质的更具体的示例（非穷尽列表）包括以下：具有一根或多根导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦除可编程存储器（EPROM或闪存/SSD存储器）、光盘、便携式高密度盘只读存储器（CD-ROM）、光存储设备、磁存储设备、RAID或前述任何适当的组合。在该文档的背景下，计算机可读存储介质可以为能够包含或存储由指令执行系统、装置或设备使用或与指令执行系统、装置或设备相结合的程序的任何有形介质。
[0078] 计算机可读信号介质可以包括其中具体实现了计算机可读程序代码的传播数据信号，例如在基带中或作为载波的一部分。这种传播信号可呈现为任意多样形式，包括但不限于电子的、电磁的、光的、或其任意适当的组合。计算机可读信号介质可以为不是计算机可读存储介质并且能够传送、传播或传输程序以便由指令执行系统、装置或设备使用或与指令执行系统、装置或设备相结合的任意计算机可读介质。
[0079] 具体实现计算机可读介质上的程序代码可以利用任何适合的介质进行传输，包括但不限于无线、有线、光纤电缆、RF等，或前述任意适当的组合。
[0080] 用于实施本发明方案的操作的计算机程序代码可以一种或多种编程语言的任意组合来编写，包括诸如Java、Smalltalk、C++等面向对象的编程语言，以及诸如“C”编程语言或类似编程语言的常规的过程编程语言。程序代码可完全在用户计算机上，部分在用户计算机上，或部分地在用户计算机上使用，部分地在远程计算机上或完全地在远程计算机或服务器上执行。在后者的方案中，远程计算机可通过任何类型的网络与用户计算机连接，包括局域网（LAN）或广域网（WAN），或可与外部计算机连接（例如，通过使用因特网服务提供商的因特网）。
[0081] 下面参照根据本发明的实施例的方法、系统和计算机程序产品的流程图和/或框图来描述本发明的方案。将理解的是，流程图和/或框图中的每个框以及流程图和/或框图中的框的组合能够通过计算机程序指令来实现。这些计算机程序指令可提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器以制成机器，使得经由计算机的处理器或其它可编程数据处理装置执行的指令生成用于实现在流程图和/或框图的一个或多
个框中指定的功能 / 动作的手段。

[0082] 这些计算机程序指令还可存储在能够引导计算机和其他可编程数据处理装置或其他设备以特定方式运作的计算机可读介质中，使得存储在计算机可读介质中的指令生成包括实现流程图和 / 或框图的一个或多个框中指定的功能 / 动作的指令的制造品。

[0083] 计算机程序指令还可以装载到计算机、其它可编程数据处理装置或其它设备上以使得在计算机、其它可编程装置或其它设备上执行一系列操作步骤以生成计算机实现的过程，使得在计算机或其它可编程装置上执行的指令提供用于实现流程图和 / 或框图的一个或多个框中指定的功能 / 动作的处理。

[0084] 现在参照图 1，图 1 为示例性的设备 (legacy)NFS 存储系统的使用率曲线图 100 的表示示例的图示，其在系统管理员实施的引退过程中具有未用尽的 (under-utilized，利用不足的) 数据存储控制器。在该示例中，管理员在 1 月份开始过程，数据存储控制器保持活跃了 9 个月，同时表 102 表示的数据存储容量存储控制器的相关使用百分比正随时间减小，直到最终控制器实际上排空了存储数据并且由系统管理员关断。

[0085] 现在参照图 2，图 2 为根据本发明的一些实施例的诸如 pNFS 存储系统的存储系统 200（任选的（同时（concurrent）检索配置系统 200）的示意图，该存储系统包括元数据服务器 (MDS) 201 和向多个同时检索客户端 203 提供存储服务的在 pNFS 中也称为数据服务器 (DS)202 的多个存储设备。任选地，元数据服务器 201 根据诸如 pNFS 协议的协议将数据记录到诸如 SAN 块级数据存储器和 NAS 文件级数据存储器的各种类型的存储设备 202 中的存取数据记录器 211 中，表示存取操作，诸如读和 / 或写操作。存取数据记录器 211 可以监控从客户端 203 接收到的多个请求 (layout request)。元数据服务器 201 可以为基于软件的服务器，或其他服务器 206 的基于硬件的服务器，并且其中例如存储服务器的一个或多个存储设备 202 可以一起寄于相同的主机上。在使用时，存储系统 200 处理数据控制请求 (例如，布局请求) 调用请求 / 布局返回请求，并且多个存储设备 202 处理数据存取请求，例如，数据写和检索返回 (retrieved) 请求。

[0086] 任选地，元数据服务器 201 包括一个或多个处理器 206，除了处理器之外，在本文还称为存储器（例如，本地闪存或 SSD 存储器）、通信设备（例如，网络接口、存储接口）以及互连单元（例如，总线、外围设备）等等。处理器 206 可以包括中央处理单元 (CPU) 并且控制系统 200 的操作。在一些实施例中，处理器 206 通过执行存储在存储器中的软件或固件来实现这些。处理器 206 可以为或者可以包括一个或多个可编程通用或专用微处理器、数字信号处理器 (DSP)、可编程控制器、专用集成电路 (ASIC)、可编程逻辑器件 (PLD) 等，或者这些设备的组合。多个元数据服务器 201 还可以并行使用。在这样的实施例中，元数据服务器 201 利用例如节点协调协议进行协调。为了简要，任意数量的元数据服务器 201 在本文称为元数据服务器 201。

[0087] 现在参照图 3A-3E，这些图为根据本发明的一个实施例的在表示状态机的流程图下运行的方法的示例性流程图，其中状态反应动作并且转移箭头涉及关于特定布局而执行的内部或外部触发。其中该状态机演示了仅仅在 pNFS 存储系统将遗留数据从一个系统存储控制器迁移到另一个，这是通过以子文件粒度级 (granularity) 来回收布局 (pNFS / 独立式 pNFS MDS) 以及将老旧数据 (虚拟化存储) 重新引导至新控制器的能力完成的。该状态机表示本发明。在本发明的一个可能的方法实施例中，与现有技术的存储管理解决方案所需
要的数级的相当长的持续时间相比，展示出可以在数小时或数天内执行整个迁移处理。而且，在所提出的实施例解决方案中，不存在丢失极少使用的客户端应用程序的风险。图3是根据本发明的一些实施例的状态机的流程图300，其所描述用于存储控制器引退的方法，其中仅在图2所示的系统200的并行存取网络文件系统的pNFS存储下运行。

[0088] 在使用时，现在参考图3A和图3B，当我们正在处理NAS型服务器引退的情况时，如流程图300中所示，典型的pNFS架构并行存取存储系统200管理员在初始阶段302决定开始其中一个系统数据存储控制器（202）的引退过程，通常是由所选控制器老化（aging）或者由于引退控制器相关的技术操作故障问题而开始引退。第一控制器引退方法步骤304与pNFS元数据系统（MDS）管理相关联，其从MDS新分配选项列表中提取与所选存储控制器相关联的卷，而不由MDS为新的文件/块/对象分配需求而使用。这将防止在引退的卷上创建新的数据以及在该过程中稍后对其进行再次定位的需要。阶段306为循环激活阶段，关于将每个所选控制器卷的数据传递到对于驻留在将要引退控制器上的每个卷的新选的控制器分配，该阶段开始对引退的控制器存储数据的内部处理。步骤308为内部第二较低级等级的子循环激活阶段，关于将每个所选控制器文件的数据传递到对于驻留在将要引退控制器上的每个文件的新选的控制器分配（allocation），该阶段开始对引退控制器存储数据的内部子处理。

[0089] 决策阶段310管理评估步骤，其分析要引退存储控制器数据内容的所选文件。具体地，310检查手头的文件是否为由客户端（203）生成的数据文件或由MDS（201）生成的特殊文件（例如，目录），这些是否存储在DS（202）上。如果为文件，则序列继续到阶段312以管理每个数据块，该数据块组合在阶段308所完成的所选文件，并且如果所选的数据块为目录，则在阶段311下系统将目录数据迁移到新选的控制器（202）中的所选卷。步骤312为内部第三较低级等级的子循环激活阶段，关于每个所选控制器数据块的数据传递到驻留在将要引退控制器所选文件上的每个数据块的新选的控制器分配，该阶段开始对引退控制器存储数据的内部子处理。在所选文件中选择具体数据块之后，步骤314处的MDS将为其自身添加标记而不接受对于所选块的新的布局请求。结果，试图使布局进入从步骤316以及直到步骤326的该特定字节范围的客户端（203）将获得重试应答。通过使用较小的数据块，MDS可以减少数据块被拒绝访问的持续时间。下一步骤316涉及MDS系统发送指令以一旦则被给予则返回布局（CB_LAYOUTRECALL）。这与相关布局副本被发送到客户端，相关布局副本为至具有或使用将要引退控制器中的布局的所有系统客户端的布局调用消息，或者可选地系统将该消息发送至所有的系统客户端。下面的步骤涉及系统本身，或者通过系统管理员手动指示系统，设置了租赁时间时间钟，租赁时间时钟限定了系统将等待与在步骤316中发出的CB_LAYOUTRECALL请求有关的所有寻址的客户端应答的最大持续时间。

[0090] 决策步骤320是通过之前的步骤316发起的，其发出请求至所有的系统客户端以检查它们是否正在使用相关匹配布局。如果由系统接收到的匹配布局反馈应答不存在，则在步骤312中所选的相关数据块通过系统在步骤324中迁移到新的卷中以便存储在由系统选择以更换老旧引退控制器的一个或多个新选的更换控制器中。可选地，如果存在来自客户端的对匹配布局应答的肯定性确认，则步骤322开始，这表示执行如在步骤318中所限定而创建的，为由318时间时钟生成的租赁时间期间内等待寻址客户端反馈应答而生成的等待延迟，直到客户端LAYOUTRETURN由系统接收到，或者租赁时间等待时间延迟过期，在此
期间未接收到 LAYOUTRETURN 客户端反馈。在该阶段步骤 324 被触发，并且相关所选数据块
由系统移除并且从旧旧控制器卷提取到另一新选的替换存储控制器上的新卷。总之，由步骤
组 314、316、318、322 和 324 所表示的旧旧控制器引退降级过程代表在本发明方法下将老
旧控制器数据传输到新选的替换存储控制器的整个提出步骤序列，均涉及驻留在所选卷
上的所选文件中的所选数据块，所述所选卷驻留在引退存储控制器上。
[0091] 步骤 326 为另一决策阶段，用于检查在引退控制器中是否存在需要迁移到新的
控制器的更多的相关数据块，如果存在另外的相关数据块，则系统返回到步骤 312 并且开
始新的块状态评估过程和迁移循环，通过执行另一步骤循环 314、316、318、322 和 324 来完
成。该循环循环重复，直到所选文件中的所有数据块从旧旧的待引退控制器迁移到新选的控
制器。当所选文件中的最后的块被检测到并迁移到新选的控制器或者多个新选的控制器
时，系统随后开始在决策步骤 328 中评估是否仍存在待从引退存储控制器迁移的新的相关
文件。如果是，则开始在转移箭头触发 329 附加循环下指示的环反馈，其中本发明的旧旧控
制器引退过程返回到步骤 308，并且对于包括在下一所选评估和存储数据迁移文件中的所
有块再次开始迁移过程。当在阶段 306 中所选的卷中的所有相关文件已经被评估并且其数
据内容从引退控制器传递到新选的存储控制器时，则系统移至决策步骤 330。
[0092] 决策步骤 330 检查在引退控制器中是否存在待评估以将其数据内容从旧旧引退
控制器传递到新选的控制器的附加卷。如果存在待检查以为其数据内容传递的附加卷，则
在指示附加循环的转移箭头触发 331 下的循环动作开始，其中处理返回到 306 以启动和再
次重复对于将要引退控制器中的整个下一评估卷的内容评估和数据传递处理。当引退控
制器中的所有卷均己由系统评估且其数据内容已经传递到新选的控制器时，决策步骤 330 在
该阶段中指示其中系统已经结束如最后阶段 336 中所述的所选引退控制器引退处理的阶
段。在该阶段中，pNFS MDS 系统考虑拆除旧旧的引退控制器并且将引退控制器停止运行
过程结束的指示发送到存储管理员。
[0093] 作为该引退处理方法的面向任选系统客户端的操作安全附加级，可以执行包含阶
段 332 和 334 的任选处理循环。该任选阶段将控制器删除通知发送到每一个系统客户端以
使它们知道所选的引退服务器不再操作并且所有其卷不存在用于其应用的相关数据。该循
环是任选的，因为在任意情况下 pNFS 系统的 MDS 服务器具有与新控制器数据内容和数据组
织有关的所有所需的更新地址数据，从而客户端能够直接并且不会存在进一步中断地存
取对于其应用所需的新的相关布局，在该阶段，所述新的相关布局均驻留在新选的以及相
关数据更新的控制器中。
[0094] 与现有技术中与引退处理有关的遗留 NFS 系统控制器的更长的持续时间相比，用
于在 pNFS 系统管理下移动整个数据内容及其从旧旧的待引退控制器传递到新选的控制器的
过程的上述方法步骤使能实现极短且高效的存储控制器老化循环。
[0095] 在使用时，参照图 3D 和图 3E，当我们处理 SAN 型服务器引退时，如图流程图 350 中
所示，典型的 pNFS 架构并行存取存储系统 200 的管理员在初始阶段 352 决定开始系统数据
存储控制器 (202) 中的一个的引退处理，典型地引退由于所选控制器老化 (aging) 或者由于
引退控制器相关联的技术操作故障问题而发起。第一控制器引退方法步骤 354 与 pNFS
元数据系统 (MDS) 管理相关联，其从 MDS 新分配的列表中提取与所选存储控制器相关联的
LUN，而不是 MDS 新分配选列表中提取与所选存储控制器相关联的 LUN。这将防止在引退 LUN 上创建
新的数据以及在处理中稍后对其进行重新定位的需要。

阶段 356 为循环激活阶段，关于将对于每个所选控制器 LUN 的数据传递到驻留在将要引退控制器上的每个 LUN 的新选的控制器分配，该阶段开始对引退控制器存储数据的内部处理。步骤 358 为内部较低级等级的子循环激活阶段，关于每个所选控制器数据块的数据传递到驻留在将要引退控制器上的每个数据块的新选的控制器分配，该阶段开始对引退控制器存储数据的内部处理。在选择具体的数据块之后，在步骤 360 中 MDS 将为其自身加标记而不接受对于所选块的新的布局请求。结果，试图布局到从步骤 362 到步骤 372 的特定字节范围的客户端 (203) 将得到重试应答。下一步骤 362 涉及 MDS 系统发送指令以便在一旦给予时返回布局 (CB_LAYOUTRECALL)。这以相关布局副本发送到客户端，相关布局副本作为至具有或使用将要引退控制器中的布局的所有系统客户端的布局调用消息，或者可选地系统将该消息发送到至所有的系统客户端。下面的步骤涉及系统本身，或者通过系统管理员预处理手动指示系统，提供了租赁时间空间，租赁空间被限定了系统将等待与在步骤 362 中发出的 CB_LAYOUTRECALL 请求有关的所有的寻址的客户端应答的最大持续时间。

决策步骤 368 是通过之前的步骤 364 发起的，其发出请求至所有的系统客户端以检查它们是否正在使用相关匹配布局。如果由系统接收到的匹配布局反馈应答不存在，则在步骤 358 中所选的数据块通过系统在步骤 370 中移动到系统所选以替换老旧引退控制器的选定替换控制器上的 LUN。可选地，如果存在来自于客户端的对匹配布局的应答的肯定性确认，则步骤 366 开始，这表示执行如在步骤 364 中所限定而创建的。为由 364 时间空间生成的租赁空间期间内等待寻址客户端反馈应答而生成的等待延迟，直到客户端 LAYOUTRETURN 由系统接收到，或者租赁空间等待时间延迟过期，在此期间未接收到 LAYOUTRETURN 客户端反馈。在该阶段步骤 370 被触发，并且相关所选数据块由系统移除并且从老旧控制器 LUN 提取到另一新选的替换存储控制器上的新的 LUN。

总之，由步骤组 360、362、364、366 和 370 所表示的老旧控制器引退降级过程代表在本发明方法下将老旧控制器数据传输到新选的替换存储控制器的整个提出的步骤序列，均涉及驻留于驻留在引退存储控制器上的所选 LUN 上的所选数据块。

步骤 372 为另一决策阶段，关于检查在引退控制器中是否存在需要迁移到新的控制器的更多的相关数据块，如果存在另外的相关数据块，则系统回退到步骤 358 并且开始新的块状态评估过程和迁移循环，通过执行另一步骤 (循环 360、362、364、366 和 370) 来完成。该循环环重复，直到所有数据块从旧的待引退控制器迁移到新选的控制器组中。当所选 LUN 中的最后的块被检测到并且迁移到新选的控制器或者多个新选的控制器时，系统随后开始在决策步骤 376 中评估。如果仍然存在待从引退存储控制器迁移的新的相关块，则其返回到步骤 358。否则，系统移至决策步骤 376。

决策步骤 376 检查在引退控制器中是否存在待评估以将其数据内容从老旧引退控制器传递到一个或多个新选的控制器的附加 LUN。如果存在待检查以为其数据内容传递的附加 LUN，则在指示附加循环的转移箭头触发 331 下的循环动作开始，其中处理返回到步骤 356 以启动和再次重复对于将要引退控制器中的数据整个下一评估 LUN 的内容评估和数据传递处理。当引退控制器中的所有 LUN 均已由系统评估且其数据内容已经传递到新选的控制器时，决策步骤 376 在该阶段中指示其中系统已经结束如最后阶段 336 所描述的所选引退控制器引退的阶段。在该阶段中，pNFS MDS 系统考虑拆除老旧的引退控制器并且将引
退控制器停止运行过程结束的通知发送到存储管理员。

【0101】 现在参考图 3C，该引退处理方法的面向任选系统客户端的操作安全附加级，
可以执行包含阶段 332 和 334 的任选处理循环。该任选阶段将控制器删除通知发送到每
个系统客户端，以使它们知道所选的引退服务器不再操作，并且所有其 LUN 不存在用于其应
用的相关数据。该循环是任选的，因为在任意情况下 pNFS 系统的 MDS 服务器具有与新控
制器数据内容和数据组织有关的所有所需的更新地址数据，从而客户端能够直接并且不会
存在进一步中断地存取在该阶段均驻留在新选的以及相关数据更新的控制器中的其应用
所需的新的相关布局。

【0102】与现有技术中与引退处理有关的遗留 NFS 系统控制器的更长的持续时间相比，用
于在 pNFS 系统管理中移动整个数据内容及其从老旧的待引退控制器传递到新选的控制器
过程的上述方法步骤使能实现极短且高效的存储控制器老化循环。

【0103】现在参考图 4，该图为在系统管理员实施引退的过程中示例性的现有技术中具有
未用尽（不足利用）数据的存储控制器的 pNFS 存储系统的使用率曲线图 400 的示意的图示。
在该实施例中，通常可以在短的持续时间内执行全部迁移，实际上为数小时至数天，结果所
有选控制器引退处理将在不到一个月内完成。从具有在 pNFS 存储下运行的数据的存储
控制器迁移数据可视为极高效率和高耗时，极短。在该情况下，当现有技术已知的引退处理相
比时，我们可以实质性地缩短控制器退出的时间段，该时间段通常由控制器中的使用容量
和网络负荷来设定，是管理员乐于承受的。在图 4 的曲线图中示出了控制器容量使用对于
时间的该极高效的耗时的处理。其中灰色条 402 表示所选控制器的数据存储容量的 pNFS
数据百分比对时间。为了开始引退处理，系统 pNFS MDS 开始极快的从老旧的待引退数据控
制器到新选的控制器的控制文件数据传递处理。该处理极其并行化且保持开启，直到最终
数据存储控制器有效地排空数据且准备由管理员停止运行。在仅 pNFS 数据存储实施例情
况所需的典型的停工期间中，控制器引退处理阶段事实上可以在几天或更少的典型程序时
间内执行。

【0104】现在参照图 5，该图为根据发明的另一实施例的在表示状态机的流程图下运行
的方法的示意图，其中状态反映动作并且转移箭头涉及内部或外部触发，内部或外部触发
针对一定的布局而执行，其中从具有不在 pNFS 存储下运行的数据的存储控制器迁移数据
可认为更难、复杂且极其耗时。在该实施例中，在引退期间的存储使用率在相同的待引退控
制器上组合了遗留非 pNFS 存储以及新的临时性 pNFS 部分数据存储空间使用。在该实施例
中，我们不缩短控制器淡出或引退的时间段，而是可选地集中于在系统管理员使老旧控制
器引退的过程中所需的整个时间段内提高老旧控制器存储容量使用率。

【0105】图 5A-5B 是描述根据发明的一些实施例的通过在诸如图 2 所描述的系统中的并行
存取网络文件系统的 pNFS 存储下运行而使包含遗留的非 pNFS 数据的存储控制器高效地引
退的方法的状态机的流程图 500。在使用时，如流程图 500 中所示，典型的 pNFS 架构并行
存取存储系统 200 管理员在初始阶段 502 决定开始一个系统数据存储控制器（202）的引退
处理，通常引退是由于所选控制器老化或者由于引退控制器相关的技术操作故障问题而发
起的。第一控制器引退方法步骤 504 与在引退处理期间内定义期望控制器使用目标参数
（Desired utilization）的存储管理员相关联。Desired utilization 是作为在数据容量
百分比中相对于控制器最大存储容量的总数据存储有效和动态存储容量的参数。Desired_
参数是通过将引退控制器的老旧的遗留有效数据存储容量与在引退期间内系统将保存在引退控制器上的新的临时性 pNFS 数据存储容量组合来一起而实现的。系统管理员还在步骤 504 中定义新的 LUN 或新的卷，以驻留在引退控制器存储空间内，其中新的 LUN 或卷被租借或租借给作为系统一部分的 pNFS MDS 服务器。新的 LUN 的选择涉及到引退控制器为 SAN 块级数据存储控制器的情况，并且新的卷的选择涉及到引退控制器为 NAS 文件级数据存储控制器的情况。

步骤 506 是这样一个步骤：它涉及到为系统设置周期性激活的监视程序以动地监控控制器数据存储使用率。通常设定用于一个月或更频繁。步骤 508 为等待下一周期性监视指示，或等待管理员请求以重新计算控制器动态变化的当前总存储有效数据存储容量或响应于系统管理员请求而取出将要引退的存储控制器的系统指示。步骤 510 为决策步骤，其中系统需要在下面的步骤 512 中启动的计算序列下重新计算控制器的当前动态变化的容量使用率，或者取出 (evict) 引退控制器并且进入阶段 502，其中控制器准备好好在系统经过处理 300 之后停止运行，或者利用控制器作为 pNFS DSC (202)。在决策步骤 510 中的重新计算选项能够周期性地启动或者通过管理员的具体重新计算请求来启动。

步骤 512 通过测量当前状态，定义为 (Periodic_utilization) 的老旧待引退控制器的动态变化的、老旧遗留的非 pNFS 数据存储容量使用来开始计算序列。下面的步骤为决策步骤 514，其中系统基于步骤 512 测量到的老旧遗留数据结果来决定：当控制器遗留数据内容达到或包含基于最大容量老旧遗留非 pNFS 数据存储容量水平的预定最终控制器引退处理初始化下的残留老旧数据内容百分比的状态时，结束控制器使用并且然后选择通往最后阶段 520 的路径 515。可选地，如果引退控制器中的老旧的非 pNFS 数据内容仍在引退控制器中的预定最大容许残留非 pNFS 数据内容以上，则系统继续下面的计算步骤 516。

根据一个实施例，系统询问管理员，如果引退控制器中的老旧的非 pNFS 数据内容仍在引退控制器中的预定最大容许残留非 pNFS 数据内容以上，则如何继续，但是减少老旧的非 pNFS 数据没有进展。在步骤 516 中，系统计算在定义为 Periodic_free_space=Desired_utilization - Periodic_utilization 的计算程序下对分配给由 pNFS MDS 管理的池的周期自由空间。步骤 516 程序的计算结果随后用于下面的步骤 518，其中系统将计算出的 Periodic_free_space 数据容量结果作为 pNFS 资源添加 (典型地作为调整大小操作) 到在步骤 504 中创建的 LUN/ 卷中。在该处理在计算出 Periodic_free_space 结果之后的下一步骤是通过关闭环 519 而返回到步骤 508 来完成的。在步骤 508 中，在监视调度时间延迟 (或异步管理员请求) 之后，如果随后测量的 Periodic_utilization 控制器数据容量使用参数仍在非 pNFS 数据水平的最小量以上或者不在，则系统开始另一评估循环。

在连续的 Periodic_utilization 计算循环序列之后系统达到足够低的 Periodic_utilization 老旧的非 pNFS 数据存储容量使用量结果时，仅仅随后系统通过阶段 514 和转移箭头触发 515 达到最后的阶段 520。在该阶段中，系统自动地检测，或者可选地，系统管理员手动地检测，引退控制器数据存储容量在该阶段只具有非显著的非 pNFS 量的存储的老旧遗留非 pNFS 量的数据留在控制器上，并且并行地大多数 pNFS 临时数据驻留在控制器上，随后在该阶段通过系统执行引退相对短的持续时间的程序 300。程序 300 结束时，控制器有效地排空可用数据并且随后停止运行，或者通过系统本身自动地或者通过系
统管理员手动地停止运行。根据一个实施例，管理员还能够决定保持控制器以其新格式活跃，100%的pNFS DS（202）。

【0110】现在参考图6，该图为本发明的另一示例性实施例的示例使用率曲线图600的图示。其中在系统管理员实施引退的过程中从未用尽数据的控制器迁移遗留数据，同时参考数据不在pNFS存储下运行的情况，因此该过程更复杂且耗时。在该情况下，我们可不缩短控制器淡出的时间段，但是可选地集中于提高在停工时间周期内的老旧控制器使用率。在该实施例的具体示例下，管理员已经在一月份开始处理，并且数据存储控制器保持活跃了9个月。并在地在该期间内，非pNFS数据存储容器和曲线图条602的存储控制器黑色区域的相关使用百分比逐渐减小，同时并行地临时租借/租赁给pNFS MDS数据存储容器和存储控制器的相关pNFS MDS数据使用百分比上升，从而继续保持存储控制器的最大数据存储容量。图6中的黑色条602代表非pNFS数据（与图1所示相似的表现），并且灰色条604表示临时租借/租赁给支持存储虚拟化的pNFS MDS的渐增的容量部分。

【0111】本实施例典型的使用曲线图600演示了，在所有该期间内，非pNFS数据存储容器和存储控制器的相关使用百分比602逐渐减小，同时临时租借/租赁给pNFS MDS数据存储容器和存储控制器容量604的相关pNFS MDS数据使用百分比上升，以确保在整个引退处理期间持续维护引退控制器最大存储使用容量所需的方式由pNFS MDS同步，直到存储控制器完全包含仅临时租借/租赁的pNFS MDS数据。在该阶段，管理员能够开始在图3的第一本发明实施例方法中描述的控制器引退处理中的短持续时间第二阶段。在该阶段，系统开始从老旧的待引退数据控制器到新选的数据控制器的快速的逐块数据传递处理，该处理保持继续直到最后数据存储控制器排空存储的数据并且准备由管理员停止运行。第二数据控制器引退处理阶段所需的附加停工时间周期通常在达到数天的级别上。

【0112】尽管已经参照有限数量的实施例描述了本发明，本领域技术人员将理解，本发明不受本文中所特别显示和描述的内容所限制。相反，本发明的范围包括本文描述的各个特征的组合和子组合，以及本领域技术人员在阅读说明书时显而易见且不是现有技术的变型例和改进。
图 1
图 2
pNFS MDS 管理员从其新分配提取与旧控制器相关联的卷，而不用于新的分配

驻留在将要引退控制器上的每一个卷

每一个文件/目录

文件

文件引导？

目录（无布局）

将数据迁移到另一控制器上的卷

每一个数据块

对于所选的块不接受新的布局请求（重试）
pNFS MDS 考虑拆除老旧的控制器并且将引退控制器停止运行过程结束的通知发送到存储管理员
图 3D
以相关布局副本（或全部）将 CB_LAYOUTRECALL 发到客户端

设定租赁时间时钟

等待直到客户端布局返回或租赁时间过期

是，确认应答

匹配布置？

无匹配布局应答

将数据迁移到另一控制器上的 LUN

下一块

块？

最后的大块

下一 LUN

LUN？

最后的 LUN

图 3E

26
图 4

pNFS 存储器上的遗留数据
存储管理员定义期望控制器使用率 (DESIRED_UTILIZATION) 并且定义新的 LUN/卷。新的 LUN/卷租借或租赁给 pNFS MDS 服务器。

设定监视器

等待周期性监视器或管理员的重新计算的请求或管理员的逐出将要引退存储控制器的请求

请求逐出将要引退的控制器

重新计算?

周期性或重新计算请求

图 5A
在期望水平下的控制器遗留数据

511

计算待分配给由 pNFS MDS 管理的池的自由空间:
PERIODIC_FREE_SPACE = DESIRED_UTILIZATION - PERIODIC_UTILIZATION

518

将 PERIODIC_FREE_SPACE 添加到新的 LUN/卷（通过调整大小）

515

系统/管理员检测到仅非显著的非 pNFS 数据量剩余在控制器上并且然后执行引退程序 300

图 5B
图 6