[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 9/44 (2006.01)
G06F 9/46 (2006.01)

[12] 发明专利说明书

专利号 ZL 200410063652.6

[45] 授权公告日 2008 年 8 月 13 日

[11] 授权公告号 CN 100410874C

[22] 申请日 2004.7.9
[21] 申请号 200410063652.6
[30] 优先权
[32] 2003. 7.10 [33] US [31] 60/486,627
[32] 2003.12.10 [33] US [31] 10/733,459
[73] 专利权人 微软公司
地址 美国华盛顿州
[72] 发明人 D·笛奥多西油 N·S·布妖尔纳
[56] 参考文献
EP1160682A1 2001.5.12
CN1078336A 1993.11.10
CN1223780A 1999.7.21
US20030079100 A 2003.4.24
CN1452737A 2003.10.29
US6192443 B1 2001.2.20
US5784615A 1998.6.21
审查员 赵 婧
[74] 专利代理机构 上海专利商标事务所有限公司
代理人 陈 斌

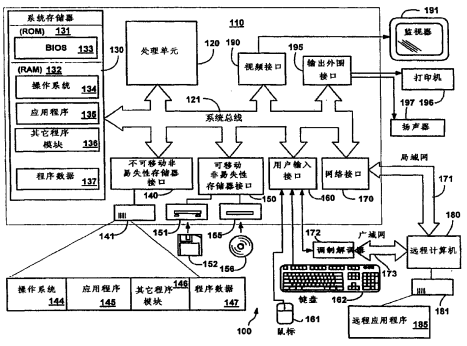
权利要求书 3 页 说明书 12 页 附图 7 页

[54] 发明名称

复制数据的方法和系统

[57] 摘要

一种用于控制哪一内容获得优先级并被复制的方法与系统。 复制品集合包括一组资源。 每一资源与资源数据和资源元数据相关联。 对于基于文件的系统，资源数据包括文件内容和属性，而资源元数据包括与在复制过程中协商同步相关的另外的属性。 将名为“围栏值”的附加域添加到与每一资源相关联的元数据上。 在同步过程中，比较第一围栏值。 具有最高围栏值的资源包括控制并复制的内容。 如果围栏值相等(并大于特定值)，则基于其它元数据确定控制资源。



1．一种复制数据的方法，包括：

接收复制品集合的第一成员的第一元数据，所述第一元数据包括第一围栏值和第一其它元数据；

将所述第一围栏值与第二围栏值相比较，其中所述第二围栏值是所述复制品集合的第二成员的第二元数据的一部分，所述第二元数据包括所述第二围栏值和第二其它元数据；

确定所述第一围栏值是否优先于所述第二围栏值；

根据所述确定改变所述第二成员的内容；以及

改变所述第二其它元数据以指示所述第二成员的内容中的变化，同时保持所述第二围栏值不变。

2．如权利要求 1 所述的方法，其特征在于，所述第二元数据储存在与所述第二成员的内容分离的存储器中。

3．如权利要求 1 所述的方法，其特征在于，改变所述第二成员的内容包括确定所述第一成员的内容和所述第二成员的内容之间的一个或多个差异。

4．如权利要求 1 所述的方法，其特征在于，所述第二成员的内容包括文件数据和文件属性。

5．如权利要求 1 所述的方法，其特征在于，所述第一元数据还包括概括第一资源的第一摘要，所述第二元数据还包括概括第二资源的第二摘要。

6．如权利要求 5 所述的方法，其特征在于，还包括比较所述第一摘要和所述第二摘要，并且如果所述第一摘要和所述第二摘要等效，则绕过改变所述第二成员的内容。

7．如权利要求 1 所述的方法，其特征在于，还包括：如果所述第一围栏值与所述第二围栏值等效，则比较所述第一其它元数据和所述第二其它元数据以确定是否应当改变所述第二成员的内容。

8．如权利要求 1 所述的方法，其特征在于，所述第一围栏值和所述第二围栏值中的每一围栏值被分配到所述复制品集合中其相应的成员的内容的一部分。

9．一种复制数据的方法，包括：

确定是否应当使用驻留在第一机器上的第一资源来更新驻留在第二机器上的第二资源，每一所述资源与元数据和内容关联，每一元数据包括只要相关联的资源

的内容改变时就被更新的一个或多个域以及一围栏值，每一围栏值表示是否应当使用其关联的资源来更新另一机器上的资源，该围栏值优先于其它元数据；

根据所述第二资源的围栏值指示所述第二资源不应被传播，防止从所述第二机器传播；以及

根据所述第一资源的围栏值优先于所述第二资源的围栏值，从所述第一资源更新所述第二资源，同时保持所述第一资源的围栏值和所述第二资源的围栏值不变。

10. 如权利要求 9 所述的方法，其特征在于，还包括：如果所述第一和第二资源的围栏值相等，则基于其它元数据确定哪一机器将被更新。

11. 如权利要求 10 所述的方法，其特征在于，所述其它元数据包括指示最后一次更新相应内容的逻辑时钟。

12. 如权利要求 11 所述的方法，其特征在于，围栏值指示其相应的资源可以传播到其它机器，直到具有更高围栏值的另一资源定位于另一机器上。

13. 如权利要求 9 所述的方法，其特征在于，与所述第一资源关联的元数据储存在与其相应资源分离的数据结构中。

14. 如权利要求 13 所述的方法，其特征在于，所述数据结构被破坏或删除，所述方法还包括重建所述数据结构并递减与所述第一资源关联的围栏值。

15. 如权利要求 14 所述的方法，其特征在于，还包括多次重建所述数据结构，并在每次重建所述数据结构时递减与所述第一资源关联的围栏值。

16. 如权利要求 15 所述的方法，其特征在于，还包括：如果所述第二资源的围栏值优先于所述第一资源的围栏值，则从所述第二资源更新所述第一资源。

17. 一种用于复制数据的系统，其特征在于，包括：

具有第一组资源的第一机器；

具有第二组资源的第二机器，其中，每一机器上的每一资源与元数据和内容相关联，每一元数据包括只要相关联的资源的内容改变时就被更新的一个或多个域以及一围栏值，每一围栏值指示是否应当使用其关联的资源来与其它元数据不相关地更新另一机器上的资源，其中，所述第一和第二机器包括：

配置成传递有关两个机器都包含的资源的信息的通信机构；以及

更新机构，配置成依照以下优先级更新每一过时的资源：

确定所述机器之一上的资源的围栏值是否优先于另一机器上的相应资源的围栏值；

如果所述机器之一上的资源的围栏值更优先于另一机器上的相应资源的所述围栏值，则使用所述一个机器上的所述资源来更新所述另一机器；否则

基于数据而非所述围栏值来更新所述机器上的资源，同时保持所述围栏值不变。

18. 如权利要求 17 所述的系统，其特征在于，所述第一组资源从备份中加载，并且设置其围栏值以使所述第一组资源优先于任一其它资源组，使得与所述资源组相应的任一其它机器上的任一其它资源组从所述第一组资源更新。

19. 如权利要求 17 所述的系统，其特征在于，所述第一组资源的围栏值被标记为优先于其它机器上的相应资源，使得所述相应资源从所述第一组资源更新。

复制数据的方法和系统

技术领域

本发明一般涉及计算设备，尤其涉及资源复制系统。

背景技术

机会主义、多主复制系统允许对参与给定复制品集合的任一机器上的复制内容进行无限制的变化。使用一组冲突分解标准在复制系统的控制下协调这些潜在冲突变化，冲突分解标准为每一冲突情况定义了哪一冲突的变化优先于其它变化。过去，使用的主要冲突分解标准为变化的物理或逻辑时间，最近的变化优先于所有其它变化。

然而，当用户或应用可能希望有另外的方法来控制哪一并发的更新优先于其它更新，并且相反地，哪一更新应放弃优先级时，存在许多问题。需要一种控制哪一内容获得优先级并被复制的灵活的方法和系统。

发明内容

简单而言，本发明提供了一种用于控制哪一内容获得优先级并被复制的方法和系统。复制品集合包括一组资源。每一资源与资源数据和资源元数据关联。对文件而言，资源数据包括文件内容和属性，而资源元数据包括有关在复制过程中协商同步的另外的属性。向与每一资源关联的元数据添加名为“围栏值（fence value）”的附加域。在同步过程中，比较围栏值。具有最高围栏值的资源包括正在控制的内容并获得复制。如果围栏值相等（并大于一特定值），则基于其它元数据确定控制资源。

围栏值与对内容的本地变化不相关。即，尽管内容中的本地变化可能影响其它元数据（如，时间标记、时钟值或其它），内容中的本地变化不影响围栏值，除非另外指示。

在本发明的一个方面，资源可能具有指示该资源未围栏的围栏值。如果该资源未围栏，则这指示不应从储存该资源的机器发送该资源。当接收到竞争的资源用于同步时，未围栏的资源负于（并且在同步过程中被替代）已围栏的资源。

在本发明的另一方面，仅发送具有获胜资源的机器和具有失败内容的机器上的资源之间的差异。例如，可能发送资源元数据而不发送资源内容。作为另一示例，可能在同步过程中发送资源内容的差异。

结合附图阅读以下详细描述，将清楚其它优点。

附图说明

图 1 是可以在其中结合本发明的计算机系统的方框图；

图 2 是依照本发明的各方面包括复制资源的两机器的资源复制系统的方框图；

图 3 是依照本发明的各方面在其中两机器试图协调两者都包括的资源的系统的方框图；

图 4 示出依照本发明的各方面能够用于图 2 和 3 的机器的一些示例性资源数据和元数据；

图 5 示出依照本发明的各方面可以使用的一些示例性资源数据和元数据；

图 6 是依照本发明的各方面一般代表了可能出现的用以同步两机器之间的资源的示例性步骤的数据流图；

图 7 是依照本发明的各方面一般代表了可能出现的用以执行非授权备份的示例性步骤的数据流图；

图 8 所示是依照本发明的各方面被配置以在资源复制系统中运行的示例机器的方框图。

具体实施方式

示例性操作环境

图 1 所示是适合在其中实现本发明的计算系统环境 100 的示例。计算系统环境 100 仅为合适计算环境的一个示例，并非对本发明的使用或功能的范围的任何限制。也不应将计算环境 100 解释为对示例性操作环境 100 中所示的组件的任一或其组合具有任何依赖或需求。

本发明适用于众多其他通用或专用计算系统环境或配置。可以适合使用本发明的众所周知的计算系统、环境和/或配置包括（但不限于）个人计算机、服务器计算机、手持式或膝上设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费者电子产品、网络 PC、小型机、大型机、包括任一上述系统或设备的分布式计算环境等等。

可以采用由计算机执行的一般语境的计算机可执行指令（如程序模块）来描述本发明。一般而言，程序模块包括例程、程序、对象、组件、数据结构等等，执行特定的任务或实现特定的抽象数据类型。本发明也可以在分布式计算环境中实践，其中，任务由通过通信网络连接的远程处理设备来执行。在分布式计算环境中，程序模块可以位于本地和远程计算机存储媒质中，包括存储器存储设备。

参考图 1，用于实现本发明的示例性系统包括以计算机 110 形式的通用计算设备。计算机 110 的组件可包括（但不限于）处理单元 120、系统存储器 130 以及将包括系统存储器的各类系统组件耦合至处理单元 120 的系统总线 121。系统总线 121 可以是若干种总线结构中的任一种，包括存储器总线或存储器控制器、外围总线以及使用任一多种总线体系结构的本地总线。作为示例而非局限，这类体系结构包括工业标准体系结构（ISA）总线、微通道结构（MCA）总线、增强 ISA（EISA）总线、视频电子标准协会（VESA）本地总线以及外围部件互连（PCI）总线，也称为 Mezzanine 总线。

计算机 110 通常包括多种计算机可读媒质。计算机可读媒质可以是可由计算机 110 访问的任一可用媒质，并包括易失和非易失媒质以及可移动和不可移动媒质。作为示例而非局限，计算机可读媒质可包括计算机存储媒质和通信媒质。计算机存储媒质包括易失和非易失、可移动和不可移动媒质，以用于储存信息的任一方法和技术来实现，信息如计算机可读指令、数据结构、程序模块或其它数据。计算机存储媒质包括但不限于，RAM、ROM、EEPROM、闪存或其它存储器技术、CD-ROM、数字多功能盘（DVD）或其它光盘存储装置、磁存储设备磁带、磁盘存储装置或可以用来储存所期望的信息并可由计算机 110 访问的任一其它媒质。通信媒质通常在诸如载波或其它传输机制的已调制数据信号中包含计算机可读指令、数据结构、程序模块或其它数据，并包括任一信息传送媒质。术语“已调制数据信号”指采用对信号中的信息进行编码的方式来设定或改变其一个或多个特征的信号。作为示例而非局限，通信媒质包括有线媒质，如有线网络或直接连线连接，以及无线媒质，如声学、RF、红外和其它无线媒质。上述的任一组合也应当包括在计算机可读媒质的范围之内。

系统存储器 130 包括易失和/非易失存储器形式的计算机存储媒质，如只读存储器（ROM）131 和随机存取存储器（RAM）132。基本输入/输出系统 133（BIOS)包含如在启动时帮助在计算机 110 内的元件之间传输信息的基本例程，通常储存在 ROM 131 中。RAM 132 通常包含处理单元 120 直接可访问和/或当前

正在操作的数据和/或程序模块。作为示例而非局限，图 1 说明了操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137。

计算机 110 也可包括其它可移动/不可移动、易失/非易失计算机存储媒质。仅作示例，图 1 说明了对不可移动、非易失磁媒质进行读写的硬盘驱动器 140、对可移动、非易失磁盘 152 进行读写的磁盘驱动器 151 以及对可移动、非易失光盘 156，如 CD ROM 或其它光媒质进行读写的光盘驱动器 155。能够在示例性操作环境中使用的其它可移动/不可移动、易失/非易失计算机存储媒质包括但不限于，磁带盒、闪存卡、数字多功能盘、数字视频带、固态 RAM、固态 ROM 等等。硬盘驱动器 141 通常通过不可移动存储器接口，如接口 140 连接至系统总线 121，磁盘驱动器 151 和光盘驱动器 155 通常通过可移动存储器接口，如接口 150 连接至系统总线 121。

以上描述并在图 1 中说明的驱动器及其关联的计算机存储媒质为计算机 110 提供了计算机可读指令、数据结构、程序模块和其它数据的存储。在图 1 中，例如，将硬盘驱动器 141 描述成是储存操作系统 144、应用程序 145、其它程序模块 146 以及程序数据 147。注意，这些组件可以与操作系统 134、应用程序 135、其它程序模块 136 以及程序数据 137 相同，也可以与它们不同。这里对操作系统 144、应用程序 145、其它程序模块 146 以及程序数据 147 给予不同的数字，来说明至少它们是不同的副本。用户可以通过输入设备，如键盘 162 和指向设备 161（通常指鼠标、轨迹球或触摸板）向计算机 110 输入命令和信息。其它输入设备（未示出）可包括麦克风、操纵杆、游戏板、圆盘式卫星天线、扫描仪、手持式 PC 的触敏屏幕或其它书写板等等。这些和其它输入设备通常通过耦合至系统总线的用户输入接口 160 连接至处理单元 120，但是也可以通过其它接口和总线结构（如并行端口、游戏端口或通用串行总线（USB）来连接。监视器 191 或其它类型的显示设备也通过接口，如视频接口 190 连接至系统总线 121。除监视器之外，计算机经常包括其它外围输出设备，如扬声器 197 和打印机 196，可通过输出外围接口 190 来连接。

计算机 110 可以在使用到一个或多个远程计算机，如远程计算机 180 的逻辑连接的网络化环境中运行。远程计算机 180 可以是个人计算机、服务器、路由器、网络 PC、对等设备或其它公用网络节点，并通常包括上述与计算机 110 相关的元件的许多或全部，尽管在图 1 中仅示出了存储器存储设备 181。图 1 所示的逻辑连接包括局域网（LAN）171 和广域网（WAN）173，但也可包括其它网络。这类网络环境常见于办公室、企业范围计算机网络、内联网和因特网。

当在 LAN 环境中使用时，计算机 110 通过网络接口或适配器 170 连接至 LAN 171。当在 WAN 环境中使用时，计算机 110 通常包括调制解调器 172 或其它装置，用于通过 WAN 173，如因特网建立通信。调制解调器 172 可以是内置的或外置的，通过用户输入接口 160 或其它合适的机制连接至系统总线 121。在网络化环境中，所描述的与计算机 110 相关的程序模块或其部分可以储存在远程存储器存储设备 181 中。可以理解，示出的网络连接是示例性的，也可以使用在计算机之间建立通信链路的其它装置。

资源复制中的控制更新

图 2 所示是依照本发明的各方面包括复制资源的两机器的资源复制系统的结构图。机器 201 和 202 复制资源 A 并且并发地将其从具有内容 x 更新为具有内容 y、u 或 z。内容 x、y、u 和 z 可能对应于例如资源 A 的不同版本的序列号和时钟。

术语“机器”不仅限于物理机器。相反，单个物理机器可包括多个虚拟机器。如这里所使用的，从一个机器复制到另一机器指将同一复制品集合的一个或多个成员从一个虚拟或物理机器复制到另一虚拟或物理机器。单个物理机器可包括同一复制品集合的多个成员。由此，复制复制品集合的成员可涉及同步包括同一复制品集合的两个或多个成员的单个物理机器的成员。

复制系统通常对每一资源保留两个相关的数据集合：资源数据和资源元数据。在基于文件系统中命名的文件的包括数据存储的复制系统中，资源数据可包括文件内容以及与文件内容关联地储存在文件系统中的任一文件属性。文件属性可包括存取控制表（ACL）、创建/修改次数以及与文件关联的其它数据。在不基于文件系统中的命名的文件的包括数据存储的复制系统（如，其资源储存在数据库或基于对象的数据存储中的系统）中，储存对数据存储合适的资源数据。在整篇文档中，经常使用基于文件系统中的文件的复制系统来说明，但是应当理解，在不脱离本发明的精神或范围的情况下可以使用能够储存内容的任一数据存储。

资源元数据包括与在复制过程中协商同步相关的一组另外的属性。对于每一资源，资源元数据可包括全局唯一标识符（GUID）、资源是否已被删除、版本序列号以及变化的原作者、反映出现变化的时间的时钟值以及其它域，如概括资源数据值并可包括对资源内容的签名的摘要。例如，摘要可用于复制同步

过程中的快速比较以绕过数据传输。如果目的地机器上的资源与源机器上的内容同步（如，由摘要指示），可以通过仅发送资源元数据而不发送资源数据其本身来最小化网络额外开销。发送资源元数据，使得目的地机器可以在其随后的复制行动中反映包括在源机器上的元数据。例如，这使目的地机器能够在随后的复制行动中成为源机器。在不脱离本发明的精神或范围的情况下，资源元数据可以与资源数据一起储存，或者可以与其分开储存。

一般而言，在广域分布式系统中，在非常粒度的级别上假定时钟同步是不可行的。这阻碍了资源系统使用全局时钟来对并发的更新和创建确定胜者。复制器通常使用在元数据上对分布式内容标记的逻辑分布式时钟。当更新内容时，逻辑时钟递增，与之形成对比的是逻辑时钟被变化时刻的物理本地时钟所覆盖。因此，逻辑时钟考虑因果性：对同一内容的更新使用永久递增的时钟值来标记。例如，图 2 的时钟值 A:x 可以为四（“4”）。复制系统然后确保与 A:y 和 A:z 关联的时钟值大于 4。这些时钟值之间的关系是任意的。值可能相等也可能不相等，因为它们是独立赋值的（如，机器 201 向时钟值赋值 A:y，而机器 202 向时钟值赋值 A:z）。

在正常的复制同步中，可以基于最后一个书写者胜出的冲突分解策略来使用时钟值来确定冲突的胜者。具有最高时钟值的数据可以指示比具有较小时钟值的复制数据更近的数据。最后一个书写者胜出的策略与逻辑时钟一致，因为它们都保持了因果性。

依照本发明的一个方面，使用名为“围栏值”的数值域来扩充资源元数据。围栏值可以分配到每一资源或其部分。在冲突分解过程中，依照一组下文定义的规则，结合其它元数据来使用围栏值。

在本发明的一个实施例中，依照规则，围栏值被初始化为 0 或 1。0 值表示不应将该资源发送到另一机器或令其对另一机器可见（通过复制机制）。1 值表示该资源可以被复制并且可令其对其它机器可见。围栏值为 0 的资源可以被认为是从资源，而围栏值为 1 的资源可以被认为是主资源。

如果在两机器间的复制行动过程中，机器之一（如机器 201）上的资源（如资源 A）比另一机器（如机器 202）上的同一资源具有更高的围栏值，则具有更高的围栏值的资源复制到具有较低围栏值资源的机器上。换言之，具有更高围栏值的资源胜出（并被复制），无论现存什么其它资源元数据。

在复制已在两机器上存在但不同的资源时，可以使用试图尽可能少地传播

数据来同步资源的机制。例如，作为对发送与资源相关联的所有数据的替代，复制机制可以确定在发送机器上的资源中的哪些数据与在接收机器上的资源中的数据不同，并发送一个或多个差异或增量来更新接收机器上的资源。

如果在复制行动过程中，围栏值相同并且大于 0，则哪一资源胜出（并被复制）取决于与每一资源相关联的其它资源元数据。换言之，当围栏值相等时，复制依照与复制相关联的普通规则进行。

由此，围栏值提供了对冲突分解进程的较佳粒度控制，因为它们在冲突分解中采用了优先级。即，当比较两资源的元数据时，具有最高围栏值的资源占先。仅当围栏相等时比较其它属性，如逻辑时钟。

围栏值可能具有与逻辑时钟类似的特性，如在一个实现方式中它们只能增加，或复位为 0。例如，当用户或进程命令复制系统增加围栏值时，围栏值增加。有时候这被称为“对资源进行围栏（fence）”或简称为“围栏”。对围栏值的增加与对资源数据的更新不相关。围栏增量由复制变为可见（如，在元数据中发送）。由于围栏不是一种频繁的操作，可以使用壁钟（wall-clock）的整数表示来将围栏增加至（当前围栏值加 1）与（当前壁钟时间）的最大值。

当用户或进程指示复制系统将围栏值复位为 0 时，围栏值被复位为 0。有时候这也被称为“对资源进行解围栏（unfence）”，或简称为“解围栏”。被解围栏的从模式资源（即，其围栏值被复位为 0 的资源）可能无法复制到除保留它们的机器之外的其它机器上。这防止了从模式资源变为外部可见。因此，围栏值复位不被复制变为可见。

除围栏和解围栏之外，围栏值在复制过程中保持恒定。具体说来，当更新资源数据或当逻辑时钟改变时，正围栏值不改变。当对从机器更新内容时，0 值的围栏值应当也不改变。

应当理解，在一个机器上对资源进行围栏允许强迫对所述资源的复制，而不管对该复制品集合中其它成员上的资源的并发更新。还应当理解，对资源进行解围栏允许强迫任一其它复制的资源优先于本地（未围栏）资源并防止了未围栏的资源被复制出去。

可以通过对所选择的资源进行围栏的复制系统应用编程接口（API）在现存的资源上设置围栏。此外，可以通过展现当资源变为对复制系统可见时对匹配具体参数（如资源名称和属性）的资源进行围栏的 API 以在未来的资源上设置围栏。也可以通过指定粘附性（stickiness）来提供对未来围栏和解围栏的

控制：创建间接相关的资源，如位于已围栏/未围栏目录下的资源，可能继承基于在父资源上设置的策略的围栏值。

解围栏也提供了一种控制复制对方的主或从行为的粒度化方式。在选择资源上设置围栏有效地令机器成为该版本的内容的主机器（直到该内容首次被更新）。对选择的资源进行解围栏令机器成为对于与选择的资源相关的内容的从机器。

应当理解，可以在许多领域内使用围栏和解围栏，包括备份恢复、带外复制、升级到复制器的更新版本以及提供管理员控制。

包括未围栏资源（如，其围栏值为 0 的资源）的机器可以对在从机器上本地更新的资源选择将未围栏的资源围栏值改变为已围栏资源值（如，1）。例如，可以完成这一过程来使更新可见。

图 3 所示是依照本发明的各方面在其中两机器试图协调两者都包括的资源的系统的结构图。在图 3 中，每一机器具有该资源的其自己的版本，并且该资源具有同一名称或标识符。在复制的结尾，目的是拥有一个或者具有内容 x 或 y 的已复制资源。在图 3 所示的系统中，机器 301 的内容 x 战胜机器 302 的内容 y。例如，当机器 301 上的资源 A 的围栏值较高，或者通过对其它资源的元数据进行比较时，会出现这一情况。

图 4 所示是依照本发明的各方面可以用于图 2 和图 3 的 A:x 的一些示例性资源数据和元数据。资源元数据包括围栏值 1、时钟值、GUID、作出变化的复制品成员以及摘要。资源数据包括资源名称、数据其自身、创建时间、修改时间以及资源数据的其它属性。

图 5 所示是依照本发明的各方面可以使用的一些示例性资源数据和元数据，与图 3 所示的同步相关。机器 501 上的内容（如，x）战胜机器 502 上的内容（如，y），并因此可以复制到机器 502，因为与 x 关联的该资源的围栏值（即，1056603359）比与 y 关联的围栏值（即，1）大。注意，在常规复制方法中，y 将战胜 x，因为它具有更大的更新时钟时间。

图 6 所示是依照本发明的各方面一般代表可能出现来同步两机器之间的资源的示例性步骤的数据流图。进程在块 605 开始。

在块 610，确定围栏值的至少一个是否指示了该资源应当被传播。可以将两围栏值设置为未围栏状态（如，0）。在这一情况下，该资源不应被传播。如果围栏值的其中之一被设置为已围栏值（如，1 或更大），如则需要，应当

传播该资源来同步两资源。

在块 615，如果该资源应当被传播，则进程分支到块 620；否则进程分支到块 635。在块 620，确定围栏值是否相等。如果是这样，则进程分支到块 625，使用元数据来确定如何传播该资源。在块 627，该资源或其部分（如，元数据、围栏值、内容差异等等）被从在块 625 确定的机器传播。

如果在块 620 围栏值不相等，则进程分支到块 630。在块 630，该资源或其一部分（如，元数据、围栏值、内容差异等等）被从具有较高的资源围栏值的机器传播到具有较低的资源围栏值的机器。在块 635，进程结束。

以下是可在其中使用本发明的一些示例性情形。

非授权备份恢复：当数据被破坏或丢失时，管理复制资源系统的成员的用户可以通过资源复制系统擦除该成员并要求获取与该成员相关联的所有数据。当与将成员连接到资源复制系统的链路的带宽相比较，成员相对较大时，这一行动的过程可能消耗较长时间或较多成本。然而，使用本发明的一个方面，用户可以从备份中恢复成员的资源。用户然后可以对备份中的资源进行解围栏（即，将围栏值设为 0），并允许资源复制系统更新过时的资源。在这一情况下，恢复的内容仅担当“内容高速缓存”的角色，给予适当的逻辑，可以由复制器用来避免通过慢链路（即，“通过线”）传输内容，由此，在受需要更新的元数据和资源限制的备份恢复之后保持了初始的同步通信量。在同步之后，任何围栏值为 0 的剩余资源可以被删除或可以将其围栏值设为 1 来允许它们被复制。

带外复制：复制品集合的新成员可能潜在地包括大量数据。为方便更快和/或较不昂贵的传输，可以通过对用户具有较低成本和/或更快服务的信道来发送新成员。例如，可以将新成员复制到硬盘或烧录到 CD-ROM 或 DVD-ROM 上，并连夜运送。在远程站点上将新成员复制到系统时，它们可能如上文关于非授权备份恢复所描述的那样被解围栏。对新成员进行解围栏避免了将复制到远程站点上的内容传输到其它站点。

授权备份恢复：用户可能希望从备份中恢复并将恢复的内容传播到复制资源系统中的所有系统。为完成这一过程，用户可以将资源恢复到系统，然后对用户希望传播的资源进行围栏。

特别情形：用户可能期望强迫复制一组资源的特定版本，并优先于复制品集合中任一其它冲突的内容。这可以通过对该组资源进行围栏来完成。

初始同步：初始同步是带外复制的一个特殊情况。在设立或升级复制系统时，用户可以将一个机器指定为主机器（即，包含将要复制到其它机器的内容），并将其它机器指定为从机器（即，从主机器接收内容）。为完成这一过程，用户可以对主机器上的资源进行围栏从而具有初始围栏值 1，并以围栏值 0 对从机器上的现有资源进行解围栏。在这一初始化之后出现在从机器上的资源可以用 1 来围栏，使其能够被复制。由于未复制未围栏的从资源，当该资源负于来自主机器的名称冲突的资源时，不需要对这一资源同步另外的元数据。

图 7 是依照本发明的各方面一般代表可能出现来执行非授权备份恢复的示例性步骤的数据流图。成员中的数据被删除或被破坏（块 710）。管理员从备份恢复资源（块 715）。管理员将恢复的数据标记为未围栏（块 720）。同步出现，如块 725 所示。管理员删除剩余资源或以围栏值 1 标记剩余资源（块 730），进程结束（块 735）。

提供图 7 作为可能在上述非授权备份中出现的步骤的一个示例。应当理解，上文中也足够详细地描述了其它应用以简单地减少到数据流图。也可以理解，可以在不脱离本发明的精神或范围的情况下对图 7 的步骤或上述应用作出许多其它变化。

图 8 所示是依照本发明的各方面被配置以在资源复制系统中运行的机器的结构图。机器 805 包括更新机制 810、资源 822 以及通信机制 840。

更新机制 810 包括围栏比较器逻辑 815，用于比较围栏值并确定资源是否应当被从机器 805 传播或资源是否应当对其它机器可见。围栏比较器逻辑 815 也可以用于确定如果发生围栏操作或资源元数据 830 的破坏或删除以及随后的重建（如下文更详细的描述）时，应当如何更新围栏值。

其它元数据冲突分解逻辑 820 包括用于确定当资源的围栏值相等时哪一资源胜出（并应当被传播）的逻辑。在应用其它元数据冲突分解逻辑 820 时，更新机制 810 可以访问资源元数据 830 和/或资源数据 825 中的数据来确定资源（或其部分）是否应当被传播到另一机器或从另一机器接收。

资源 822 包括资源数据 825 和资源元数据 830。资源数据和资源元数据之间的区别如上文参照图 2 所描述的那样。尽管在同一框中示出，资源数据 825 可能储存在相对于资源元数据 830 分离的存储中。

通信机制 840 允许更新机制 810 与其它机器上的其它更新机制（未示出）进行通信。更新机制一起确定应当同步哪一资源以及同步应当如何出现。通信

机制 840 可以是网络接口或适配器 170、调制解调器 172 或任一其它装置，用于如结合图 1 所述的那样建立通信。

可以理解，在不脱离本发明的精神和范围的情况下实现图 8 所示的机器的其它变异。

在本发明的一个实施例中，有三个初始围栏值：-1、0 和 1。-1 扮演上述 0 的角色（即，未围栏值），0 可以代表只要没有具有 1 或更高的围栏值的其它资源就可以复制该资源。这一扩展的一种使用是处理复制元数据的丢失。例如，资源复制器可以在对于用来储存复制数据的存储分离的存储中保存元数据。由此，该元数据和该资源数据（即，内容）可能不相关地失败。在重建过程中，丢失其元数据的机器可以初始地使用值-1 来对其资源进行围栏，因为它不知道该资源是否已被复制或者刚被本地创建（并且因此不被复制品集合中的其它机器所知）。过早地将围栏值从-1 改变为 1 将导致将旧内容重新引入回网络中，作为来自更新的内容，但是断开连接的机器可能被改写。另一方面，将围栏从-1 改变为 0，可以允许重新向上引入这一内容，直到某一其它参与方确定它为旧。注意，如果仅期望正整数，则替代地可以使用 0、1 和 2（或一些其它编号模式），对上述每一值移位。

本发明的再一实施例使用-∞和 0 作为可能的初始围栏值。在功能上，围栏值-∞代替了 0（即，为未围栏值），而 0 代替了 1（即，指示该资源应当被复制）。另外，可以使用其它负围栏值（如，-1、-2、-3、…、-n）来指示可以被复制的资源。这类围栏值将负于以更高的值围栏的资源。这一一般化的一种使用包括新鲜资源的可用性比恢复过程中的带宽使用更重要的情形。每一机器可能被分配在初始资源上使用的一个不同的围栏值（如，-1、-2、…、-n）。该顺序给予使用围栏值-1 的领导机器优先级，然而来自所有机器的非冲突内容立即可用。

这一表示的另一应用包括每次当机器恢复或重建其元数据作为对元数据的破坏或丢失的响应时确定围栏值。这一应用的启发值是每次当机器重建其元数据时，储存在机器上的资源数据可能变少并较不可靠和/或较不新。通过这一机制，较不频繁地重建其元数据的机器上的内容被认为是更可靠的，并且通过围栏值战胜较频繁地重建其元数据的机器上的内容。

应当理解，上述围栏值是可以在任一多种物理围栏值中实现的逻辑围栏值。例如，在不脱离本发明的精神或范围的情况下，围栏值可以物理地被表示

为无符号整数、浮点数、比特值或任一其它类型的编号模式。

从上述详细描述中可以看到，本发明提供了一种用于粒度化地控制复制信息的改进的系统和方法。尽管本发明可以具有各种修改和替代构造，然而在附图中示出并详细描述了其特定的说明实施例。然而，应当理解，这并非将本发明限制在所揭示的特定形式中，而是相反，本发明覆盖了处于本发明的精神和范围之内的所有的修改、替代构造和等效物。

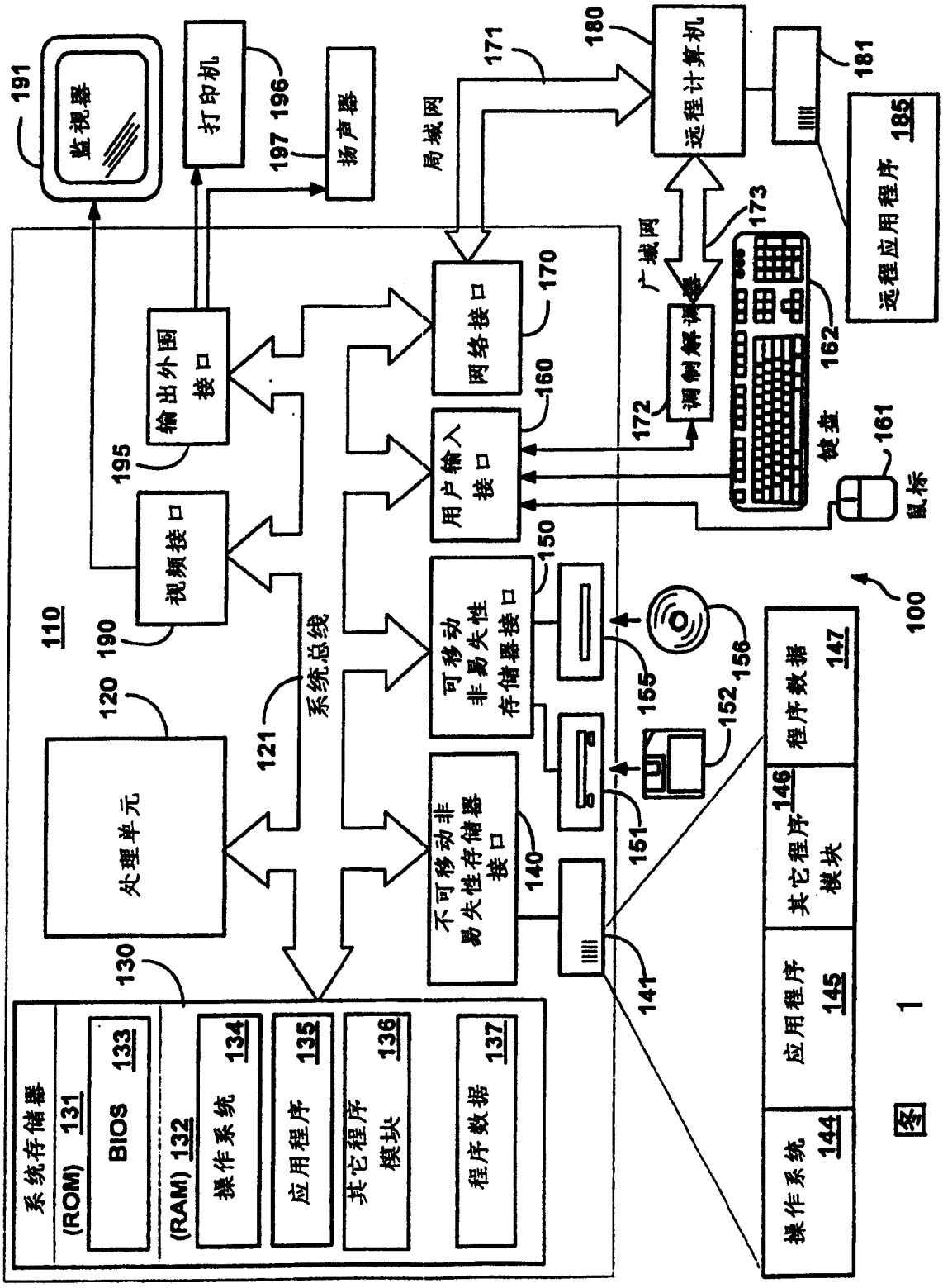

监视器
191
打印机
196
197
扬声器
局域网
171
180
远程计算机
181
185
远程应用程序
广域网
173
162
键盘
161
鼠标
输出外围接口
195
网络接口
170
160
调制解调器
172
用户输入接口
视频接口
190
110
120
处理单元
121
系统总线
150
可移动非易失性存储器接口
155
156
152
151
140
不可移动非易失性存储器接口
141
130
系统存储器
(ROM) 131
BIOS 133
(RAM) 132
操作系统 134
应用程序 135
其它程序模块 136
程序数据 137
操作系统 144
应用程序 145
其它程序模块 146
程序数据 147
100

图 1

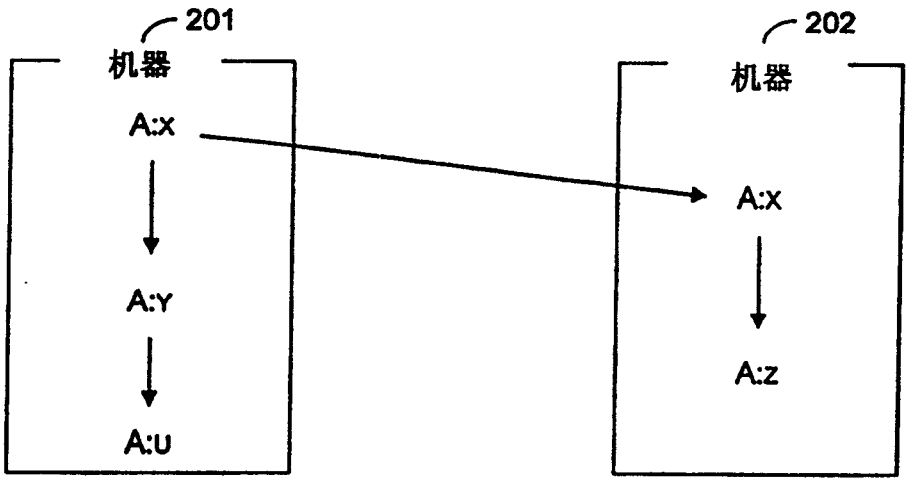


图 2

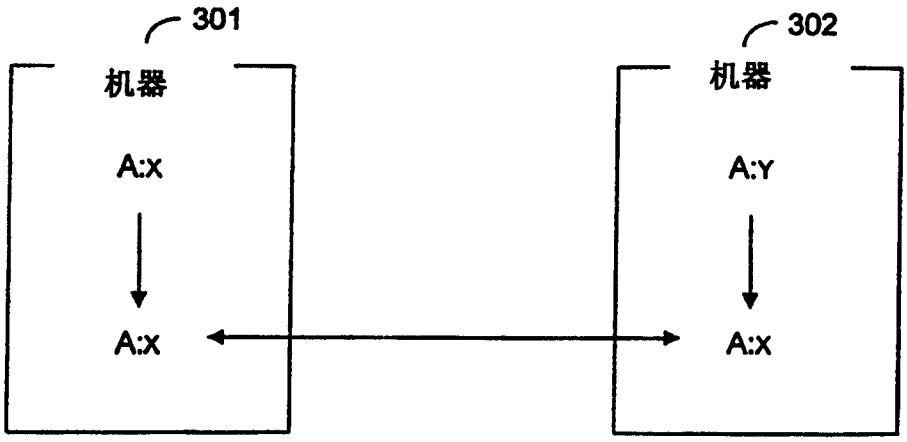


图 3

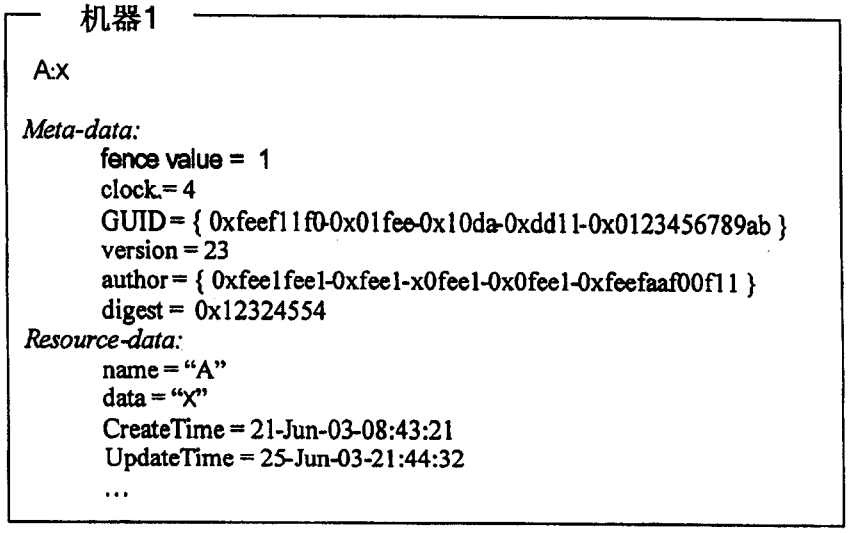

机器1
A:x
Meta-data:
fence value = 1
clock = 4
GUID = { 0xfeef11f0-0x01fee-0x10da-0xdd11-0x0123456789ab }
version = 23
author = { 0xfee1fee1-0xfee1-x0fee1-0x0fee1-0xfeefaaf00f11 }
digest = 0x12324554
Resource-data:
name = "A"
data = "X"
CreateTime = 21-Jun-03-08:43:21
UpdateTime = 25-Jun-03-21:44:32
...

图 4

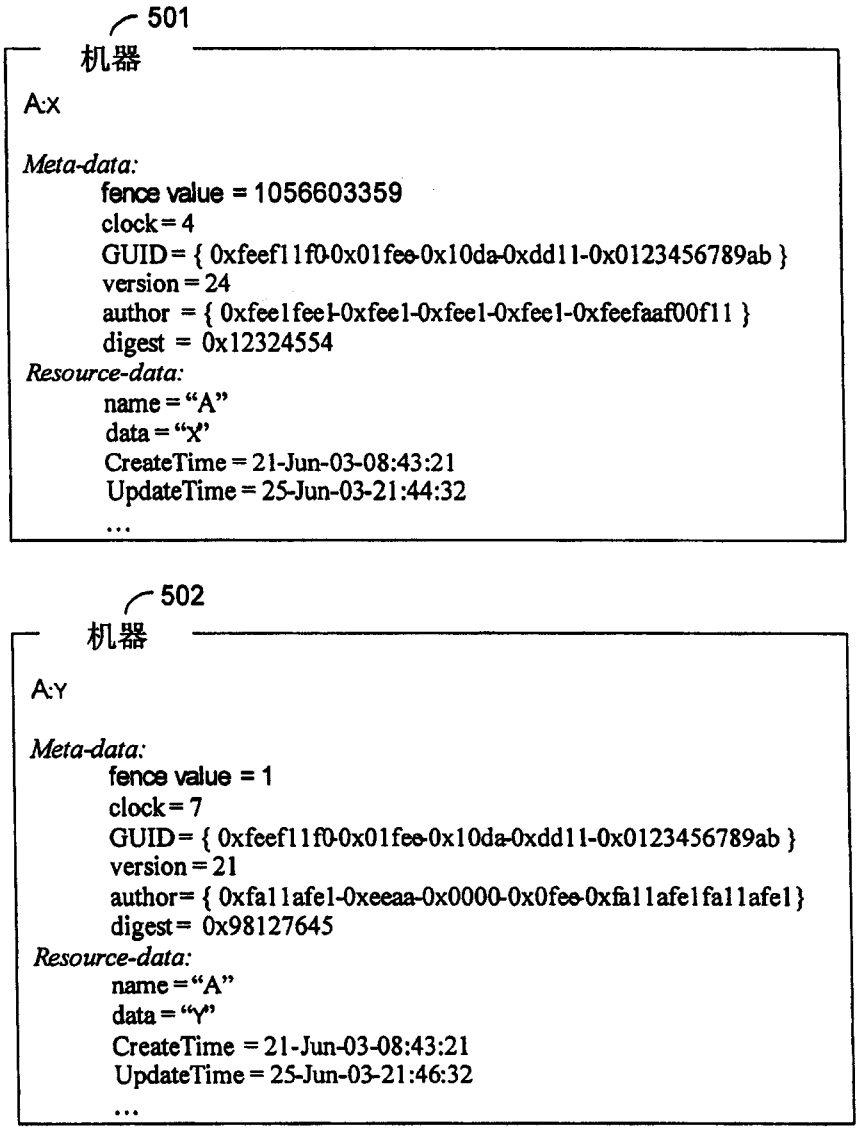

501
机器
A:X
Meta-data:
fence value = 1056603359
clock = 4
GUID = { 0xfeef11f0-0x01fee-0x10da-0xdd11-0x0123456789ab }
version = 24
author = { 0xfee1fee1-0xfee1-0xfee1-0xfee1-0xfeefaaf00f11 }
digest = 0x12324554
Resource-data:
name = "A"
data = "X"
CreateTime = 21-Jun-03-08:43:21
UpdateTime = 25-Jun-03-21:44:32
...
502
机器
A:Y
Meta-data:
fence value = 1
clock = 7
GUID = { 0xfeef11f0-0x01fee-0x10da-0xdd11-0x0123456789ab }
version = 21
author= { 0xfa11afe1-0xeeaa-0x0000-0x0fee-0xfa11afe1fa11afe1}
digest= 0x98127645
Resource-data:
name = "A"
data = "Y"
CreateTime = 21-Jun-03-08:43:21
UpdateTime = 25-Jun-03-21:46:32
...

图 5

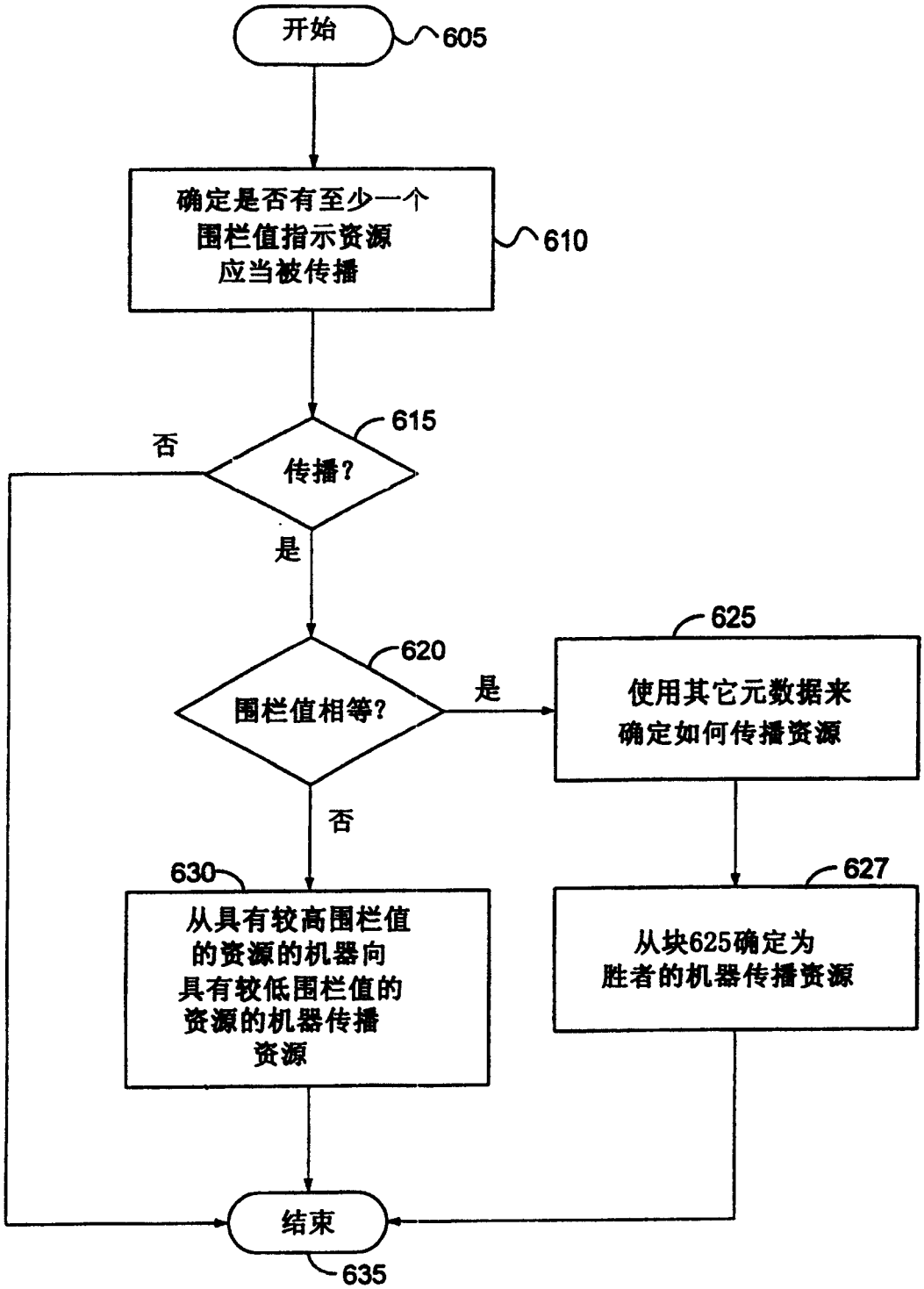

开始
605
确定是否有至少一个
围栏值指示资源
应当被传播
610
615
传播?
否
是
620
围栏值相等?
是
625
使用其它元数据来
确定如何传播资源
否
630
从具有较高围栏值
的资源的机器向
具有较低围栏值的
资源的机器传播
资源
627
从块625确定为
胜者的机器传播资源
结束
635

图 6

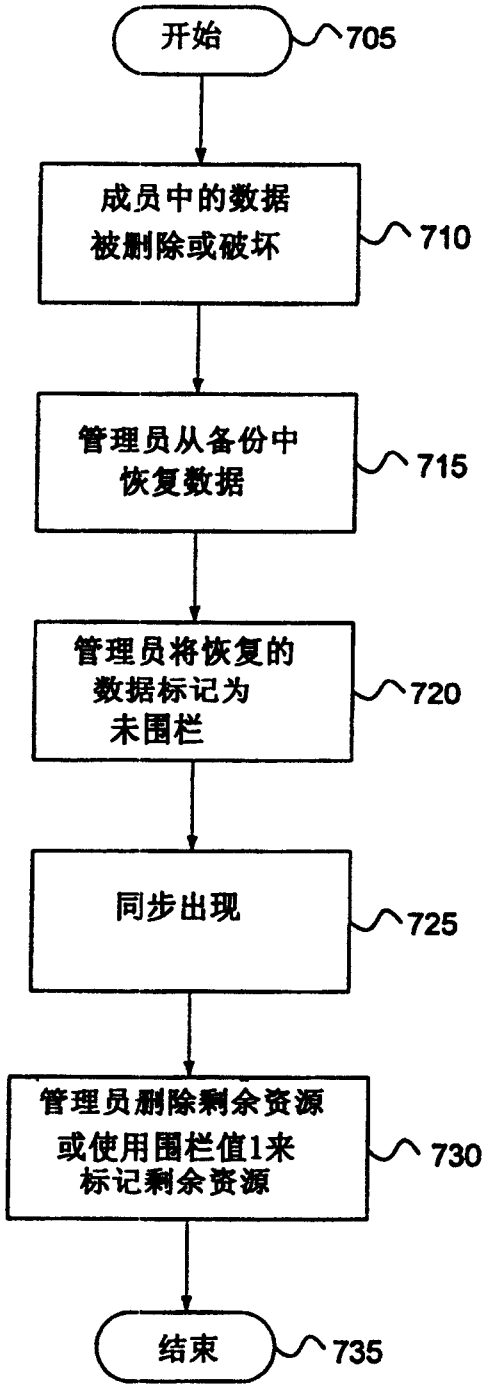

开始
705
成员中的数据
被删除或破坏
710
管理员从备份中
恢复数据
715
管理员将恢复的
数据标记为
未围栏
720
同步出现
725
管理员删除剩余资源
或使用围栏值1来
标记剩余资源
730
结束
735

图 7

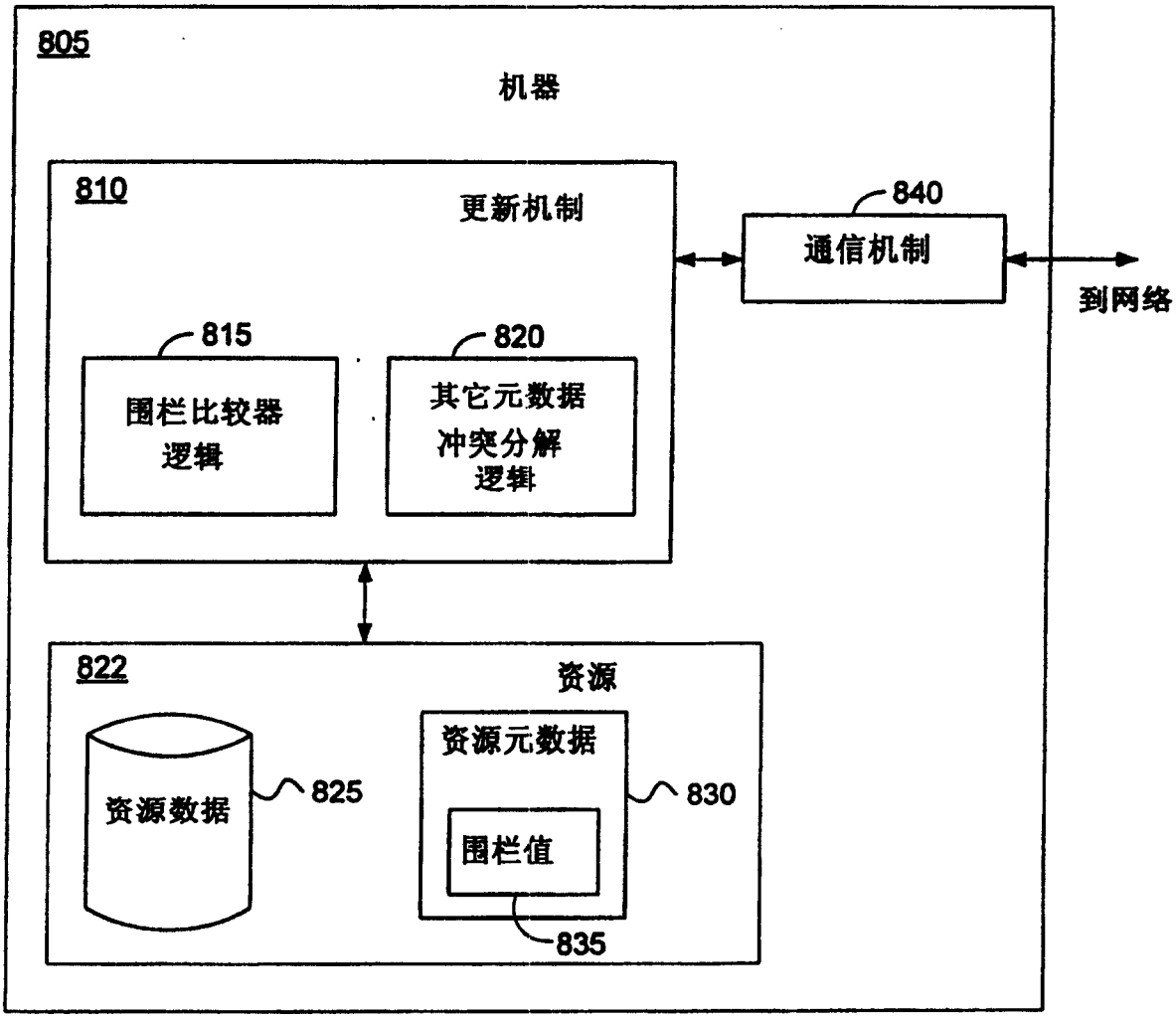

805
机器
810
更新机制
840
通信机制
到网络
815
围栏比较器
逻辑
820
其它元数据
冲突分解
逻辑
822
资源
资源数据
825
资源元数据
830
围栏值
835

图 8