



[12] 发明专利申请公开说明书

[21] 申请号 200480023059.7

[43] 公开日 2006年9月20日

[11] 公开号 CN 1836232A

[22] 申请日 2004.8.9

[21] 申请号 200480023059.7

[30] 优先权

[32] 2003.8.14 [33] US [31] 60/495,368

[32] 2003.9.3 [33] US [31] 60/500,050

[32] 2003.9.3 [33] US [31] 60/500,096

[32] 2003.11.21 [33] US [31] 10/718,747

[86] 国际申请 PCT/US2004/025805 2004.8.9

[87] 国际公布 WO2005/020102 英 2005.3.3

[85] 进入国家阶段日期 2006.2.13

[71] 申请人 甲骨文国际公司

地址 美国加利福尼亚州

[72] 发明人 本尼·苏德尔 詹姆斯·斯塔莫斯

黄力 约翰·奇米斯基

阿南德·拉克希米纳特 艾伦·唐宁

[74] 专利代理机构 北京康信知识产权代理有限责任
公司

代理人 余刚

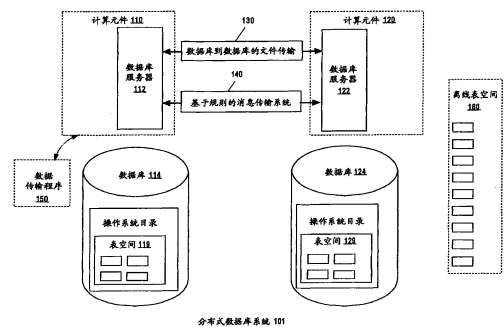
权利要求书 8 页 说明书 23 页 附图 6 页

[54] 发明名称

数据库的自动和动态提供

[57] 摘要

本发明描述的方法可以用于提供在分布式计算环境(诸如网格)中需要大量数据传输的数据库。此方法不需要 DBA 手动干预,以例如在操作系统的文件系统之间传输表空间。相反,只要网格计算系统确定需要动态地提供数据库,表空间就可以通过网格计算系统自动地和动态地被提供。另外,由于提供了表空间的副本,还可以自动提供同步机构以使表空间与其副本保持同步。



1. 一种用于在分布式数据库系统中自动供应数据的方法,所述方法包括以下步骤:

数据库服务器使表空间从第一文件系统传输到第二文件系统; 以及

在将所述表空间传输到所述第二文件系统中后, 所述数据库服务器将所述表空间导入到由所述数据库服务器管理的本地数据库中。

2. 根据权利要求1所述的方法, 其中, 所述数据库服务器使表空间被传输的步骤和所述数据库服务器导入所述表空间的步骤都响应于例程的调用而被执行。
3. 根据权利要求1所述的方法, 其中, 所述例程被写成符合数据库语言并且可以通过数据库服务器执行的代码。
4. 根据权利要求1所述的方法, 其中, 所述导入的步骤包括将所述表空间附加到所述本地数据库。
5. 根据权利要求1所述的方法, 其中, 在执行所述数据库服务器使表空间被传输的步骤之前或期间, 将所述表空间附加到另一个数据库。
6. 根据权利要求1所述的方法, 其中, 在执行所述数据库服务器使表空间被传输的步骤之前或期间, 所述表空间是离线的。

7. 根据权利要求 1 所述的方法，其中：

所述导入所述表空间的步骤包括附加所述表空间的副本，其中，所述副本与所述表空间不同；以及

所述数据库服务器提供同步机构，所述同步机构将对所述表空间所作的改变应用于所述副本。

8. 根据权利要求 7 所述的方法，其中，所述同步机构将对所述副本所作的改变应用于所述表空间。

9. 根据权利要求 7 所述的方法，其中，所述步骤还包括：

所述同步机构基于通过规则引擎产生的一组规则的评估结果，确定将对所述表空间所作的哪些改变传播到所述副本；以及

其中，所述供应同步机构的步骤包括配置所述组规则。

10. 一种用于数据库服务器提供文件副本的方法，所述方法包括以下步骤：

第一数据库服务器接收创建存储在第一操作系统的第一文件系统中的文件的副本的请求；

所述第一数据库服务器在特定操作系统的特定文件系统中创建所述副本；以及

其中，所述副本是与所述特定文件不同的文件。

11. 根据权利要求 10 所述的方法，其中：

所述第一数据库服务器接收请求的步骤包括所述第一数据库服务器接收将所述文件的副本传输到所述特定文件系统的请求；

其中，所述第一数据库服务器创建所述副本包括在所述第一数据库服务器与所述第二数据库服务器之间传输所述文件的副本；以及

在所述特定文件系统中存储所述副本。

12. 根据权利要求 11 所述的方法，其中：

所述第一文件系统相对于所述第一数据库服务器是本地的，并且相对于所述第二数据库服务器是远程的；

所述特定文件系统相对于所述第二数据库服务器是本地的，并且相对于所述第一数据库服务器是远程的；以及

其中，所述存储的步骤由所述第二数据库服务器执行。

13. 根据权利要求 12 所述的方法，其中，所述传输的步骤包括经由消息传输系统传输作为二进制文件的所述副本，所述消息传输系统在所述第一数据库服务器与所述第二数据库服务器之间传播消息。

14. 根据权利要求 11 所述的方法，其中：

所述第一文件系统相对于所述第二数据库服务器是本地的，并且相对于所述第一数据库服务器是远程的；

所述特定文件系统相对于所述第一数据库服务器是本地的，并且相对于所述第二数据库服务器是远程的；以及

其中，所述存储的步骤由所述第一数据库服务器执行。

15. 根据权利要求 10 所述的方法，其中，所述第一文件系统相对于所述第一数据库服务器是本地的，并且所述特定文件系统相对于所述第一数据库服务器是本地的。

16. 根据权利要求 10 所述的方法，其中，接收请求包括调用将识别所述文件的值作为参数的例程。
17. 根据权利要求 10 所述的方法，其中：
 - 接收请求包括通过接口接收命令；
 - 所述数据库服务器执行通过符合数据库语言的所述接口接收的命令；以及
 - 所述命令识别所述文件。
18. 一种用于在分布式数据库系统中自动地例示数据库数据的方法，所述方法包括以下步骤：
 - 数据库服务器将一组一个或多个文件从第一文件系统传输到第二文件系统；
 - 其中，所述组一个或多个文件存储用于数据库的数据；以及
 - 在将所述组一个或多个文件传输到所述第二文件系统后，所述数据库服务器供应所述数据库作为由所述数据库服务器管理的数据库。
19. 根据权利要求 18 所述的方法，其中，所述组一个或多个文件是表空间，其中，所述供应的步骤包括：
 - 将所述表空间附加至所述数据库。
20. 根据权利要求 18 所述的方法，其中，所述组一个或多个文件包括描述数据库对象的元数据和用于将数据插入到所述数据库对象中的命令，其中，所述供应的步骤包括通过执行命令将所述数据导入到所述数据库中。

21. 根据权利要求 18 所述的方法, 其中, 所述组一个或多个文件包括由恢复管理器创建的备份文件, 其中, 所述供应的步骤包括使所述恢复管理器从所述备份文件创建所述数据库。
22. 根据权利要求 21 所述的方法, 其中, 存档日志存储记录创建所述备份文件之后对所述数据库所作的改变的数据, 其中, 所述供应的步骤还包括改变所述数据库以反映记录在所述存档日志中的改变。
23. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 1 中所述的方法。
24. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 2 中所述的方法。
25. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 3 中所述的方法。
26. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 4 中所述的方法。
27. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 5 中所述的方法。

28. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 6 中所述的方法。
29. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 7 中所述的方法。
30. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 8 中所述的方法。
31. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 9 中所述的方法。
32. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 10 中所述的方法。
33. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 11 中所述的方法。
34. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 12 中所述的方法。
35. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 13 中所述的方法。

36. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 14 中所述的方法。
37. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 15 中所述的方法。
38. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 16 中所述的方法。
39. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 17 中所述的方法。
40. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 18 中所述的方法。
41. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 19 中所述的方法。
42. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 20 中所述的方法。
43. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 21 中所述的方法。

-
44. 一种计算机可读介质, 承载一个或多个指令序列, 当所述指令序列由一个或多个处理器执行时, 其使得所述一个或多个处理器执行在权利要求 22 中所述的方法。

数据库的自动和动态提供

相关申请

本申请要求由 Debashish Chatterjee 等人于 2003 年 8 月 14 日提交的题为“Computer Resource Provisioning”的第 60/495,368 号美国临时申请的优先权,其内容结合于此以供参考;本申请要求由 Benny Souder 等人于 2003 年 9 月 3 日提交的题为“Automatic And Dynamic Provisioning Of Databases”的第 60/500,050 号美国临时申请的优先权,其内容结合于此以供参考;本申请要求由 Carol Colrain 等人于 2003 年 9 月 3 日提交的题为“Service Based Workload Management and Measurement in a Distributed System”的第 60/500,096 号美国临时申请的优先权,其内容结合于此以供参考;以及本申请要求于 2003 年 11 月 21 日提交的题为“Automatic and Dynamic Provisioning of Databases”(Attorney Docket No.50277-2343)的第 10/718,747 号美国申请的优先权,其内容结合于此以供参考。

以下申请和当前申请有关:

由 Edwina Lu 等人于 2003 年 5 月 30 日提交的题为“Utilizing Rules in a Distributed Information Sharing System”的第 10/449,873 号美国申请,其内容结合于此以供参考;以及

由 Juan R. Loaiza 等人于 2003 年 1 月 28 日提交的题为“Pluggable Tablespaces For Database Systems”的第 10/353,381 号美国申请,其内容结合于此以供参考。

技术领域

本发明涉及分布式数据库系统，并且尤其涉及在数据库系统中分配资源以满足变化的工作负荷需求。

背景技术

企业正在考虑减少他们的数据处理系统的成本和增加其效率的方式。典型的企业数据处理系统为每个企业应用分配单个资源。为每个应用获取足够的资源，以处理应用的估计峰值负荷。每个应用具有不同的负荷特性；一些应用在白天繁忙；另外的一些应用在夜间繁忙；一些报表每周运行一次，而另外的一些报表每月运行一次。结果，有很多资源容量没有被利用。网格（grid）计算使得该未用的容量被利用或被删除。实际上，网格计算随时准备着彻底地改变计算的经济效果。

网格是商品计算元素的集合，其提供处理和一定程度的共享存储；网格资源被动态地分配，以满足计算需求和网格客户端的优先权。网格的实例是服务器刀片（blade）的机架（rack）。每个服务器刀片是一个兼容（inclusive）计算系统，在单个母板上具有处理器、存储器、网络连接、以及相关电子设备。典型地，服务器刀片不包括板上存储器（不同于易失性存储器），并且它们与机架内的电源、冷却系统、以及电缆一起共享存储单元（例如共享磁盘）。

网格计算能显著地降低计算的成本，扩展计算资源的可用性，并且产生更高的生产率和更高的质量。网格计算的基本思想是计算作为公用设施（utility）（类似于电网或者电话网）的概念。网格的客户端不关心它的数据在哪里或者计算在哪里执行。客户端想要的一切是使计算完成，并且当它需要的时候使信息传递到客户端。

在更高的水平上，网格计算的中心思想是计算作为公用设施。网格的客户端不必关心它的数据位于哪里，或者什么计算机元件处理请求。客户端只需要请求信息或计算，并使得它被传递 - 像需要的那样多并且无论何时需要。这类似于电气设施工作的方式；顾客不知道发电机在哪里，或者电网怎样连线。顾客只需要电并获取它。此目标是将计算做成一种公用设施 - 一种无处不在的商品。因此它具有这个名字，网格。

网格计算作为公用设施的视角当然是客户端侧的视角。从服务器侧或在后台看，网格是关于资源分配、信息共享、以及高可用性。资源分配确保需要或请求资源的所有那些获得他们需要的。当请求没有被服务时，资源不是闲置的。信息共享确保信息客户端和应用需要在需要其的地方和时候是可用的。高可用性确保所有数据和计算必须总在那里 - 正如公用设施公司必须总提供电力一样。

用于数据库的网格计算

能够从网格计算获益的计算机技术的一个领域是数据库技术。网格能支持多个数据库并根据需要动态地分配资源，以支持每个数据库上的负荷。当数据库的负荷增加时，更多的资源被分配用于该数据库。例如，在企业网络上，数据库通过运行在网格上的一个服务器刀片上的数据库服务器提供服务。从数据库请求数据的用户数量增加。响应于对数据库需求的这种增加，在一个或多个其它服务器刀片上供应另一个数据库服务器。

数据库网格的供应

此术语供应指提供并配置提供服务需要的计算资源和数据。关于数据库服务器，供应包括配置服务器刀片以运行数据库服务器，

以及配置数据库服务器以管理数据库。关于数据库，供应包括配置数据库服务器以管理对数据库的访问。

供应数据或数据库的过程在此被称为数据供应。在网格中供应数据库可能需要克隆（clone，复制）所有或部分数据库，然后供应一个新的数据库服务器以管理克隆或将克隆结合到另一个已经被已经运行的数据库服务器管理的数据库中。

数据库的数据供应可包括在文件系统和/或数据库之间的大量数据传送。不幸地是，用于数据库供应的大量数据传送的技术需要人工干预，因此不能在网格计算需要时被用于自动地和动态地供应数据。

使用用于大量数据传输技术的数据供应方法的实例是可传输表空间（tablespace）方法。表空间是用于为数据库对象（例如关系表）存储数据的存储容器（例如文件）的集合。在此方法中，表空间被从“源数据库”导出并导入到“目标数据库”中。这能够允许使用用于复制文件的操作系统应用程序复制表空间的文件，这比其他的用于在数据库之间的大量数据传输技术运行快得多。这样的其他技术包括执行查询和插入语句。

为了传输表空间，人力数据库管理者（“DBA”）执行手动步骤。首先，通过附加表空间，表空间必须被导入到目标表空间中。关于表空间、数据库、和数据库服务器，术语“附加（attach）”指配置数据库和/或数据库服务器，以使表空间中的数据库对象被结合到数据库中，并且表空间用于为数据库存储数据。配置数据库以附加表空间包括修改数据库元数据，以使它将表空间和数据库对象限定为数据库的一部分。可以使用包括由 DBA 执行的手动步骤的多种技术来修改数据库元数据。DBA 能运行在可以在源数据库系统上使用的公共设施，其中，源数据库系统可以被执行以将元数据导出到“元

数据堆栈文件”中，并且 DBA 可以运行在目标数据库系统上的应用程序，以从元数据转储文件构建元数据。可选地，元数据可以与被传输的数据一起包括在表空间中，并且目标数据库将从包括在表空间中的元数据重新构建元数据。DBA 也能在目标数据库系统上手动重新构建元数据。

通过从原始源数据库创建表空间的单独副本以及将其附加到目标数据库，可以将表空间传输到数据库。尽管进行了复制，但是对表空间的操作应当被限制为只读操作。DBA 发送命令以指示管理数据库的数据库服务器，将对表空间执行的数据库操限制为只读操作。一旦复制完成，DBA 可以发送命令以指示数据库服务器可以执行修改操作。

在此使用的术语“副本”既指源数据又指源数据的复制。例如，源文件的副本可以是源文件本身，或者是另一个文件，另一个文件是能够使用例如容易获得的副本应用程序（例如用于创建数据文件的副本的操作系统应用程序）生成的副本。

传输的表空间的副本也可以与数据库分离。关于特殊的表空间和数据库以及数据库服务器，术语“分离”指配置数据库和/或数据库服务器，以使表空间不再用于为数据库存储数据。配置数据库以分离表空间包括，通过例如除去将表空间限定为源数据库系统的一部分的元数据或通过设置标志以指示表空间不再被使用，来在源数据库系统中改变数据库元数据。通过 DBA 运行应用程序或通过手动编辑源数据库元数据来执行此步骤。

源数据库的表空间被存储在文件系统的“源目录”中，并且目标数据库系统的表空间被存储在文件系统的“目标目录”中。源目录和目标目录可以在相同的文件系统内或不同计算机系统上的不同文件系统内。在任何情况下，DBA 均需要用操作系统应用程序

来传输表空间。这需要 DBA 在目标目录的计算机系统中具有操作系统帐户。DBA 登陆到计算机系统并运行应用程序，以将表从源目录传输到目标目录。如果目标目录是另一个计算机系统的另一个文件系统，DBA 能使用 FTP 来传输文件（即使用遵从文件传输协议的公共设施）。为了使用 FTP，DBA 需要在目标目录的计算机系统上的操作系统帐户以登陆到计算机系统并传输表空间文件。

如上所述，传统的用于供应数据库的大量数据传输技术需要在人工 DBA 方面手动干预。因为网格计算需要数据供应自动和动态地执行，要求大量数据库传输的数据库供应对于网格计算机是不顺从的。明显地，需要自动化的大量数据库传输，其适于网格内的动态数据供应。

此部分描述的方法是能够实施的方法，但是不一定是以前构想或实施过的方法。因此，除非另外指出，不应该认为此部分描述的任何方法仅由于包含现有技术就等同于现有技术。

附图说明

通过举例说明本发明，而不用于限制本发明，在附图中用相同的标号标记相似的部件，附图中：

图 1 是描述本发明的实施例可以在其上执行的分布式数据库系统的框图。

图 2 是根据本发明的实施例的用于自动地和动态地供应表空间的程序的框图。

图 3 是描述根据本发明的实施例的用于自动地和动态地供应表空间的过程的流程图。

图4是描述根据本发明的实施例的用于使可以被自动地和动态地供应的表空间同步的机构的框图。

图5是根据本发明的实施例的用于自动地和动态地供应表空间和表空间同步机构的流程图。

图6是描述可以被用于执行本发明的实施例的计算机系统的框图。

具体实施方式

描述了数据库的自动和动态供应的方法和装置。在下面的描述中，为了解释的目的，为了提供对本发明的彻底理解，阐述了许多具体的描述。然而，很明显，没有这些具体描述本发明也可以实现。在其它情况下，为了避免对本发明不必要的混淆，以框图的形式示出公知的结构和装置。

此处描述的方法可以用于要求在分布式计算机环境（例如网格）中传输大量数据的数据库供应。此方法不要求DBA手动干预，例如，在操作系统的文件系统之间传输表空间。相反，无论何时网格计算系统确定需要动态地供应数据库，表空间可以通过网格计算系统自动地和动态地供应。另外，由于提供了表空间的副本，也可以自动地供应同步机构以保持表空间与其副本同步。

典型的分布式数据库系统

图1示出了分布式数据库系统101，其可以用于执行本发明的实施例。分布式数据库系统101包括数据库服务器112和数据库服务器122。数据库服务器112管理对数据库114的访问，数据库服务器122管理对数据库124的访问。数据库服务器（例如数据库服务器112和122）是一组集成软件组件与计算资源的分配的结合，

例如存储器和用于在处理器上执行这组集成软件组件的过程，其中，软件和计算资源的结合用于管理数据库。在数据库管理的其它功能中，数据库服务器控制并使得对数据库的访问容易实现，通过数据库客户端来处理请求以访问数据库。数据库服务器的数据库客户端可以包括其它数据库服务器。数据库 **110** 是数据库对象的集合。数据库对象包括任何形式的结构化数据。结构化数据是根据限定结构的元数据描述而结构化数据。结构化数据包括关系表、对象表、对象关系表、和根据可扩展标记语言（“XML”）结构化的数据体，例如 XML 文档。

表空间 **119** 和 **129** 是“在线（online，也称联机）”的，因为它们中的每个均被定义为数据库的一部分。离线（offline，也称脱机）表空间（例如离线表空间 **180**）是被分离的表空间，也就是说，没有被限定为数据库的一部分。此处描述的方法可以用于提供在线表空间和离线表空间的副本。

相对于数据库服务器 **112**，数据库 **114** 被称作本地数据库，因为数据库服务器 **112** 不需要另一个数据库服务器来访问代表数据库服务器 **112** 的数据库 **114**。然而，为了访问数据库 **124**，数据库服务器 **112** 必须从数据库服务器 **122** 请求访问，数据库服务器 **122** 代表数据库服务器 **112** 访问数据。因此，数据库 **124** 不是数据库服务器 **112** 的本地数据库。

如上所述，表空间是存储容器的集合。根据本发明的实施例，存储容器是由操作系统限定和管理的文件。操作系统是控制和管理计算机元件的资源的软件。在计算机元件上执行的其它软件（此处指“应用程序”）通过在操作系统的控制下运行的过程来执行。数据库服务器的软件组件是应用程序的实例。执行应用程序的过程对资源的访问也由操作系统控制。应用程序通过调用操作系统例行程序和应用程序来访问资源。

由操作系统限定和管理的文件被存储在永久性存储器中，例如磁盘存储器，并且被组织成包括文件和其它目录的层级目录。直接由计算机元件上的操作系统控制的文件系统是指关于操作系统的本地文件系统。文件系统文件是指关于操作系统的本地文件。如果访问文件系统不需要另一个操作系统访问代表特殊操作系统的文件系统，则文件系统被该特殊操作系统直接控制。从而，在共享磁盘系统中的操作系统可以被不只一个操作系统控制。

控制应用程序或过程的计算元件上的操作系统在此称作关于该应用程序或过程（process）的本地操作系统。同样地，应用程序或过程被称作关于该操作系统的本地应用程序或本地过程。

使在操作系统的控制下执行的过程与由负责注册帐户的操作系统功能建立的操作系统帐户相关联。为用户初始化过程要求用户使用用于登录用户的操作系统功能在帐户下登录。

当操作系统控制计算元件的资源时，控制可以与应用程序共享，应用程序与操作系统共享控制并且从属于操作系统。例如，数据库服务器 112 包括多个与操作系统帐户关联的“数据库服务器过程”。与该操作系统帐户关联的过程（其包括多个数据库服务器过程）通过计算机元件 110 的操作系统授予对于操作系统的本地文件系统中的目录的独占读写特权。由数据库服务器 112 管理数据库服务器过程对目录中的数据的访问。与其它操作系统帐户关联的过程不可以访问这些目录。这样，数据库服务器 112 基于这些目录和在这些目录内的文件共享联合控制和从属控制。

操作系统提供各种功能（function）和应用程序，其允许以特定方式管理和处理资源。例如，计算机元件 110 的操作系统可以将磁盘驱动器看作独立（或便宜的）磁盘的冗余阵列。同样地，数据库服务器能提供类似的能力。

数据库传输过程 **150** 包括可以被数据库服务器调用以执行需要供应表空间的特定组步骤的例程。程序可以执行一些或者所有需要供应表空间的步骤，包括由 DBA 常规和手动执行的步骤。可以调用这些程序以自动地和动态地供应数据库。

数据传输机构

数据库服务器 **112** 和 **122** 使用各种数据传输机构(“传输机构”)在彼此之间传输数据。这些传输机构包括从数据库服务器到数据库服务器的文件传输机构 **130** (“DB 文件传输机构”)和基于规则的消息传输 (messaging) 系统 **140**。为了说明的目的，在图 1 中使用与那些用于表示数据库服务器 **112** 和 **122** 的框分离的和不同的框描述 DB 文件传输机构 **130** 和基于规则的消息传输系统 **140**。然而，这些机构可以包括集成在数据库服务器 **112** 和 **122** 中任意一个内的软件组件，并且其参与沿着在数据库服务器之间的互连 (例如网络、总线) 传输数据。实际上，这些传输机构可以使用在数据库 **114** 和 **124** 内的队列和表，或者在数据库服务器 **112** 和 **122** 控制下的文件，并且可以使用来自数据库传输程序 **150** 的程序。

DB 文件传输机构 **130** 是用于在数据库服务器 **112** 和 **122** 之间传输文件的传输机构。文件可以被从数据库服务器 **112** 的本地文件系统中的—个或多个目录传输到数据库服务器 **122** 的本地文件系统的—个或多个目录。数据库服务器的本地文件系统是控制数据库服务器的操作系统的本地文件系统。从而，数据库服务器 **112** 的本地文件系统是计算机元件 **110** 的操作系统的本地文件系统。根据本发明的实施例，使用类似于 FTP 的协议传输文件。

DB 文件传输机构 **130** 在以下方面不同于传统的文件传输机构。它是数据库服务器组件，(1) 只以文件的形式传输数据，以及 (2) 在数据库服务器之间传输文件，其或者从本地文件系统中恢复文

件，一接收到文件，就将其存储在本地文件系统中，并且（3）响应于传输文件的请求，以此方式传输文件，其中，请求指定要传输的特定的文件和源和目的地的位置（例如目录）。传输的文件可以是二进制文件或者文本文件。DB 文件传输机构 **130** 也执行字符集转换，转换所传输的文件的字符集格式。

很清楚，尽管传输机构已经在文件系统目录之间传输了文件，但这种机构不包括作为数据库服务器的集成组件的软件组件。而且，数据库服务器 **112** 和 **122**（不同于传统的数据库服务器）被配置为识别指定文件在目录之间传输的命令。这种命令可以被用户经由命令行接口输入，命令行接口也接收符合数据库语言（例如 SQL）的查询。指定文件传输的命令指定文件名，包括文件位置的目录路径、以及文件名和文件所传输到的目录路径。

基于规则的消息传输系统

基于规则的消息传输系统 **140** 在数据库服务器 **112** 与数据库服务器 **122** 之间传送消息。消息包含关于事件的信息，事件例如数据的创建或修改。消息用于将发生在一个数据库服务器的事件传播（propagate）到另一个数据服务器。然后，其它数据库服务器可以将消息传输到又一个数据库服务器。

消息传输系统的一般用途是复制（replicate）数据。对数据库服务器上的数据库对象所作的 DML 改变被传播到保持数据库对象的至少一部分的副本的另一个数据库服务器。基于规则的消息传输系统 **140** 可以用于传输广泛类型的数据，包括文件。

从一个消息传输系统到另一个消息传输系统所需的消息流可以不同。各种类型的消息传输系统为用户提供在网络中的节点之间配置消息流的能力。一类消息传输系统是基于规则的消息传输系

统，例如基于规则的消息传输系统 **140**，其允许用户指定控制消息流的规则。

规则指定条件和动作，如果条件满足，则执行动作。一般地，规则遵守规则语言，规则语言类似计算机语言。使用规则的消息传输系统通过能被规则引用（reference，也可以称参考）的变量或属性揭露关于事件的信息。规则中的条件可以使用布尔表达式表示，布尔表达式引用变量和属性。规则可以用于选择哪个消息的哪个事件被发送到其它节点，以及对从另一个节点接收的消息如何处理。

数据库传输程序

图 2 更详细地描述了根据本发明实施例的数据库传输程序 **150**。数据库传输程序 **150** 包括文件传输程序 **210** 和可传输的表空间程序 **220**。文件传输程序 **210** 包括涉及在数据库服务器之间传输文件的例程，而表空间传输程序 **220** 包括涉及传输表空间的程序。这些程序可以由数据库服务器或用户通过用户界面（例如，命令行界面）调用。数据库服务器的调用可以指定一个或多个参数，并且可以返回一个或多个值（例如函数调用值或参数返回值）。这些参数用于指定例如分离什么表空间或传输哪个文件。

根据本发明的实施例，数据库传输程序 **150** 被用 C 和 PL/SQL™ 的结合编写。PL/SQL 是可以从 Oracle™ 公司获得的过程数据库语言。然而，本发明不限于以特定的计算机语言编写的数据库传输程序。

文件传输程序 **210** 包括程序获取（get）**212**、放置（put）**214**、和副本 **216**，如下所述。

获取 **212** 这个程序，使“目标”数据库服务器从“源”数据库服务器请求源数据库服务器本地的文件。源数据库服务器将请求的文件传输到目标数据库服务器。可以使用 DB 文件传输机构 **130** 传输请求和文件。

放置 **214** 这个程序使源数据库服务器联系目标数据库服务器，以在目标数据库服务器的本地文件系统中创建来自源数据库服务器的本地文件系统的文件的副本。源数据库服务器将请求的文件传输到目标数据库服务器。可以使用 DB 文件传输机构 **130** 传输文件。此程序可以被基于规则的消息传输系统 **140** 使用，以经由 DB 文件传输机构 **130** 传播文件。

副本 **216** 这个程序使数据库服务器产生到本地文件系统的文件的本地副本。

数据库传输程序 **150** 包括分离 **222**、克隆 **224**、附加 **226**，拉 (pull) **28**、和推 (push) **229**，如下所述。

分离 **222** 这个程序使数据库服务器使一组表空间只读，以从数据库分离表空间，并返回表空间中的文件 (“表空间文件”) 的名称。通过只兑现 (honor) 和处理从表空间读取数据的请求，以及防止对表空间中的数据的 DML 改变，数据库服务器使一组表空间只读。该组表空间的表空间元数据被导出到单独的文件，称作 “元数据转储文件”。

表空间文件和元数据转储文件一起形成 “可传输表空间包”。可以使用文件传输程序 **210** 传输可传输表空间包。以此方式分离的可传输表空间包也能被 “重新附加” 到它被从其中分离的数据库。

克隆 224 这个程序在不分离表空间的情况下复制表空间。特别地，程序使数据库服务器使一组表空间只读，以将它们的表空间文件复制到可以由参数指定的另一组表空间文件，然后使该组表空间读-写（如果在调用程序前读-写）。此程序也将用于该组表空间的元数据导出到元数据转储文件。程序返回新表空间文件和元数据转储文件的名称。

附加 226 这个程序使数据库服务器附加在可传输集合包中的一组表空间。

拉 228 这个程序使得数据库服务器从远程数据库服务器的远程数据库（不在数据库服务器本地的一个数据库）复制一组表空间，并将表空间附加到服务器的本地数据库中。从而，程序完全地供应表空间。图 3 示出了此程序。

参考图 3，在步骤 310 中，使在远程服务器的表空间只读。在步骤 320 中，数据库服务器用元数据导入/导出应用程序获取描述表的元数据。这种应用程序允许客户端和数据库服务器连接到另一个数据库服务器并获取描述数据库对象的元数据，包括描述表和表空间以及在表空间内的数据库对象的元数据。在步骤 330 中，数据库服务器使用获取 212 以获取表空间的副本，并将它们存储在本地文件系统中。在步骤 340 中，通过使用在步骤 320 中获得的表空间元数据附加表空间，将表空间导入到本地数据库中。在步骤 350 中，使在远程服务器的表空间读-写。

推 229 这个程序使源数据库服务器复制一组表空间，并将其导入到远程数据库服务器的本地数据库中。通过经由基于规则的消息传输系统 140 将消息发送到远程数据库服务器，可以执行此程序，消息指定表空间到“拉”。然后，远程数据库服务器调用拉 228，以供应本地数据库中的表空间。另一个执行此程序的方法用于源数据

库服务器使用放置 **214** 将表空间的副本传输到远程数据库，以调用元数据导入/导出应用程序来将描述表空间的表空间元数据传输到远程数据库服务器，并指示远程数据库服务器使用表空间元数据附加这些副本。

表空间同步机构的自动供应

一旦表空间的副本被供应到另一个数据库中，基于规则的消息传输系统 **140** 可以用于使表空间和它的副本保持同步。这样做，必须供应基于规则的消息传输系统 **140** 的各种元件。

图 4 是示出被供应以使表空间与其供应的副本同步的元件的框图。图 5 是描述用于自动地供应表空间副本和使表空间副本与表空间同步所需的基于规则的消息传输系统元件的过程的流程图。这些元件和过程用分布式数据库系统 **101** 示出。

图 4 更详细地示出根据本发明的实施例的基于规则的消息传输系统 **140**。参考图 4，示出了包括捕获过程 **413** 的数据库服务器 **112**，捕获过程 **413** 捕获由数据库服务器 **112** 在变化日志 **417** 中记录的事件（例如 DML 改变和 DDL 改变）。变化日志 **417** 包括重做和/或撤消记录。捕获过程 **413** 将反映那些事件的消息排队到消息队列 **418** 中。传播过程 **414** 将消息从队列 **418** 传播到消息队列 **422**。消息队列 **422** 是分级（staging）区，用于将被数据库服务器 **122** 应用到数据库 **124** 的消息。

响应于来自客户端的请求以评估在规则 **452** 或 **454** 中的规则集，规则引擎 **451** 和 **453** 分别执行规则 **452** 和 **454**，然后将评估规则集的评估结果返回到客户端。规则引擎 **451** 的客户端包括捕获过程 **413** 和传播过程 **414**。规则引擎 **453** 的客户端包括应用过程 **423**。捕获过程 **413**、传播过程 **414**、以及应用过程 **423** 使用由规则引擎

451 和 **453** 提供的规则评估结果，以确定怎样处理事件和消息。这通过将执行规则集的请求传输到规则引擎而实现。响应于这样的请求，规则引擎 **451** 和 **453** 执行一个或多个规则集。基于规则的消息传输系统在“Utilizing Rules in Distributed Information Sharing and Utilizing Rules in a Distributed Information Sharing System”中更详细描述。

通过使用图 5 中描述的过程将表空间 **419** (图 4) 的副本供应到数据库 **124** 作为表空间 **419'** 示出了该过程。数据库 **114** 被称作源数据库，因为它包括复制的表空间，并且数据库 **124** 被作为目标数据库，因为它是副本所附加到的数据库。数据库服务器 **112** 和 **122** 被分别称作源数据库服务器和目标数据库服务器，因为它们分别是源数据库 **114** 和目标数据库 **124** 的本地数据库服务器。

参考图 5，在步骤 510 中，源数据库服务器 **112** 被配置为记录对表空间 **419** 的 DML 改变（即对存储在表空间 **419** 中的数据库对象的改变）的日志，以运行捕获过程 **413** 以捕获对表空间 **419** 的改变，并运行传播过程 **414** 以将对表空间 **419** 的改变传播到目标数据库服务器 **122**。

在步骤 520 中，规则 **452** 被配置为使捕获过程 **413** 捕获对表空间 **419** 的改变，并将改变传播到目标数据库服务器 **122** 和消息队列 **422**。

在步骤 530 中，通过调用拉程序 **228**，在目标数据库服务器 **122** 上供应表空间 **419** 的克隆，表空间 **419'**。

在步骤 540 中，目标数据库服务器 **122** 被配置为运行应用过程，以应用来自消息队列 **422** 的改变。

在步骤 550 中，规则 **454** 被配置为使应用过程 **423** 将对表空间 **419** 的改变应用到表空间 **419'**（即将对表空间 **419** 内的数据库对象的改变应用到在表空间 **419'** 内的数据库对象）。

因此在图 4 中描述的过程可用于自动地和动态地供应表空间，此过程可以被作为程序执行，程序被调用以将表空间和基于规则的消息传输元件供应到特定的数据库或数据库服务器。可选地，可生成脚本，并且以后执行它以供应表空间和/或同步机构。

在图 4 中描述的过程供应同步机构，它是单向的，就是说，对表空间的改变只能从一个表空间传播到另一个表空间而反过来不行。然而，本发明的一个实施例不受如此限制。双向的同步机构也可以被自动地供应。例如，通过配置捕获、传播、和应用过程、以及以类似于所描述的将改变从表空间 **419** 传播到表空间 **419'** 的方式的规则，规则和消息流过程可以被配置为将改变从表空间 **419'** 传播到表空间 **419**。

本发明的实施例的另外特征

虽然通过供应在线表空间已经示出了供应表空间的方法，但是本发明并不限于此。此处描述的方法可以用于自动地和动态地供应离线表空间的副本。

例如，离线表空间 **180** 可以包括关于季度财务结果的信息。表空间中的数据只在每季度的短暂时间内季度地被处理和报告。在占用资源（例如，存储能力）的位置不保持表空间在线，信息可以保持离线并被季度地供应和“不供应”，释放资源和允许资源用于其他服务。

虽然此处描述的方法以表空间的形式供应数据，本发明不限于供应被表空间识别的数据。本发明可以用于自动供应一列表、大纲（schema）和数据库，以及保持它们既单向又双向使它们保持同步的同步机构。供应一列表、大纲和数据库为顾客提供更多的灵活性和合理的方式以自动地供应数据并保持其同步。

在分成多种拓扑布局的分布式数据库系统中，可以供应同步机构。通过消息传输系统互连的数据库服务器网络可以通过节点的定向图表示。边缘连接节点，每个边缘代表从“源”节点到另一个相邻的“目的”节点的消息流。对于给定的节点，多个边缘可以从该节点发出或终止于该节点。另外，可以有循环，表示消息从源节点沿着可能包括一个或多个其它节点的路径流回到源节点。在可以用各种类型的定向图（例如非循环的图或完全连通图）表示的拓扑布局中，可以自动地供应数据和/或同步机构，那里在相邻的节点之间的变化可以被双向同步。

用于自动地和动态地供应数据的方法并不限于供应数据库数据。这些方法可以用于供应软件和计算机代码，例如 PL/SQL 包。数据库服务器提供代码导入/导出应用程序，用于将代码导出到“代码转储文件”，并用于从代码转储文件导入代码。例如为了自动地供应代码，数据库服务器能调用代码导入/导出应用程序以从远程数据库服务器导出代码，并将代码存储在本地文件系统中的转储文件中，然后调用代码导出/导入应用程序将代码导入到本地数据库或代码储存库中。代码的自动供应能用于远程数据库服务器上的远程作业调度。数据库服务器能将作业或者任务授权给远程数据库服务器，并且通过自动地供应代码和数据，提供需要的代码和/或数据以执行作业或任务。

其它的基于文件的数据库供应方法的用途

通过将可传输的表空间用作例示 (instantiate) 数据库的方法示出了本发明的实施例, 也就是说生成数据库的副本并供应它。然而, 本发明并不限于此。本发明的实施例可以用其它的技术达到此目的。

例如, 可以通过使用命令产生方法例示数据库。在命令产生方法中, 数据库服务器创建包括数据库对象的描述 (即元数据) 和数据库语言命令的文件, 数据库语言命令将数据插入到数据库对象中。这种文件在此称作导入/导出文件。导入/导出文件可使用例如导出应用程序创建, 导出应用程序用于通过创建描述数据库对象的元数据和数据库对象中插入行的 SQL 插入命令, 为数据库中选定的数据库对象导出数据库数据。为了供应数据库数据, 数据库服务器执行导入应用程序, 导入应用程序用于通过读取数据库对象的描述, 确保数据库对象被限定在目标数据库 (如果需要, 限定它们) 中, 然后执行 SQL 插入命令来导入数据。SQL 插入命令的执行将数据一行一行地插入到目标数据库中, 这需要比当插入比数据阈值大的数据时附加表空间更多的工作。

这种方法对于在各种情况下供应目标数据库可能有利。其中一种情况是当目标数据库已经为要供应的数据限定数据库对象并且创建数据库对象的限定的开销因此没有发生时, 要供应的数据量小且一行一行插入的开销因此无意义。

另一种情况是, 当在表空间方法中一组要例示的数据库对象在不需要例示其它数据库对象的情况下不能被例示。表空间用于为一组特定的数据库对象存储数据。在一些表空间的实施例中, 当使用表空间以例示数据库对象时, 在该组中的所有数据库对象必须被例

示。然而，在命令产生方法中，有可能在这组表空间乃至从其它的表空间中只为选定的子集数据库对象创建导入/导出文件，赋予只例示选定的数据库对象的能力，无论哪个表空间用于存储它们的数据。这样，在形成要例示的数据库对象的结合中，命令产生方法能提供更多的灵活性。

为了动态地例示数据库的副本，数据库服务器运行导出用程序以创建导入/导出文件，然后使用数据库服务器文件传输机构或消息传输系统（例如数据库服务器文件传输机构 **130** 或基于规则的消息传输系统 **140**）在数据库服务器之间传送该导入/导出文件。可选地，导入/导出文件可以存储作为库的一部分，将在后面需要的时候被传输和供应。

另一个例示数据库的方法是恢复管理器（recovery manager）方法。恢复管理器方法使用恢复管理器的能力以例示数据库。恢复管理器用于创建所有数据库文件（例如，表空间文件）的备份文件并从备份文件中恢复数据库文件。在创建备份文件的同时，恢复数据库文件例示了数据库的副本。如果对数据库的改变被存档在存档文件日志中，则存档日志可用于恢复所作的改变。实际上，数据库能被恢复到存档日志所覆盖的任何特定的时间点。

通过使用恢复管理器创建备份文件以及使用备份和恢复管理器在另一个位置创建目标数据库，可以例示源数据库。如果在创建目标数据库之前以及在源数据库改变期间备份被存储一定时间，则源数据库存档日志可用于使目标数据库成为当前数据库或与在某时间点的源数据库一致。实际上，通过使用源数据库的存储日志更新多个实例以使它们与给定的时间点一致，可以更有效地例示多个来自一组备份文件的在不同时期的实例。

硬件概述

图 6 示出可实现本发明的实施例的计算机系统 600 的方框图。计算机系统 600 包括总线 602 或用于传送信息的其它通信机构、和与总线 602 连接以处理信息的处理器 604。计算机系统 600 还包括例如随机存取存储器 (RAM) 或者其它动态存储装置等主存储器 606, 该主存储器与总线 602 连接, 用于储存信息及要由处理器 604 执行的指令。主存储器 606 还可用于在执行将被处理器 604 执行的指令期间储存临时变量或其他中间信息。计算机系统 600 还包括与总线 602 连接的只读存储器 (ROM) 608 或者其他静态存储装置, 用于储存静态信息和关于处理器 604 的指令。例如磁盘或光盘等存储装置 610 与总线 602 连接, 用于储存信息和指令。

计算机系统 600 可以经由总线 602 连接到诸如阴极射线管 (CRT) 等用于向计算机用户显示信息的显示器 612。包括字母数字键和其它键的输入装置 614 与总线 602 相连接, 用于传送信息和命令选择到处理器 604。另一种用户输入装置是光标控制器 616, 如鼠标、轨迹球、或光标方向键等, 用于传送方向信息和命令选择到处理器 604 及用于控制显示器 612 上的光标移动。该输入装置通常具有沿两个轴 (第一轴 (例如 x 轴) 和第二轴 (例如 y 轴)) 的两个自由度, 使装置可在平面中确定位置。

本发明涉及用于实现在此描述的技术的计算机系统 600 的使用。根据本发明的实施例, 响应于处理器 604 执行包含在主存储器 606 中的一个或多个指令的一个或多个序列, 通过计算机系统 600 执行这些技术。可将这些指令从诸如存储设备 610 等另一计算机可读介质读入主存储器 606。通过执行包含在主存储器 606 中的指令序列, 使处理器 604 执行此处所述的处理步骤。在可选实施例中, 可用硬连线电路取代软件指令, 或者与软件指令结合来实施该发

明。因此，本发明中的实施例不限于硬件电路和软件的任何特定组合。

这里使用的术语“计算机可读介质”是指参与提供指令给处理器 **604** 来执行的任何介质。这种介质可以采取很多形式，包括但不限于非易失性介质、易失性介质、和传输介质。非易失性介质举例来说包括例如存储装置 **610** 等光盘或磁盘。易失性介质包括例如主存储器 **606** 等动态存储器。传输介质包括同轴电缆、铜线、和光纤，包括包含总线 **602** 的导线。传输介质还可采取声波或光波的形式，例如那些在无线电波和红外线数据通信过程中产生的声波和光波。

通常的计算机可读介质举例来说包括：软盘、软磁盘 (flexible disk)、硬盘、磁带、或者任何其它磁性介质；CD-ROM、任何其它光学介质；打孔纸、纸带、或者任何带有孔图案的物理介质；RAM、PROM、EPROM、FLASH-EPROM、或者其他任何存储芯片或者卡；下面描述的载波、或者计算机可读的任何其他介质。

各种形式的计算机可读介质可参与传送一个或者多个指令的一个或多个序列给处理器 **604** 来执行。例如，该指令初始地可承载在远程计算机的磁盘中。该远程计算机能将该指令加载到其动态存储器中，然后使用调制解调器通过电话线发送指令。计算机系统 **600** 本地的调制解调器可接收电话线上的数据，且使用红外转换器将数据转换成红外信号。红外探测器可以接收红外信号携带的数据，且合适的电路可以把信息放到总线 **602** 上。总线 **602** 把数据传送到主存储器 **606** 中，处理器 **604** 从主存储器 **606** 取回并执行这些指令。由主存储器 **606** 接收的指令可随意地在处理器 **604** 执行这些指令之前或之后储存于存储装置 **610** 中。

计算机系统 **600** 还包括连接到总线 **602** 的通信接口 **618**。通信接口 **618** 提供连接到网络链路 **620** 的双向数据通信，该网络链路与

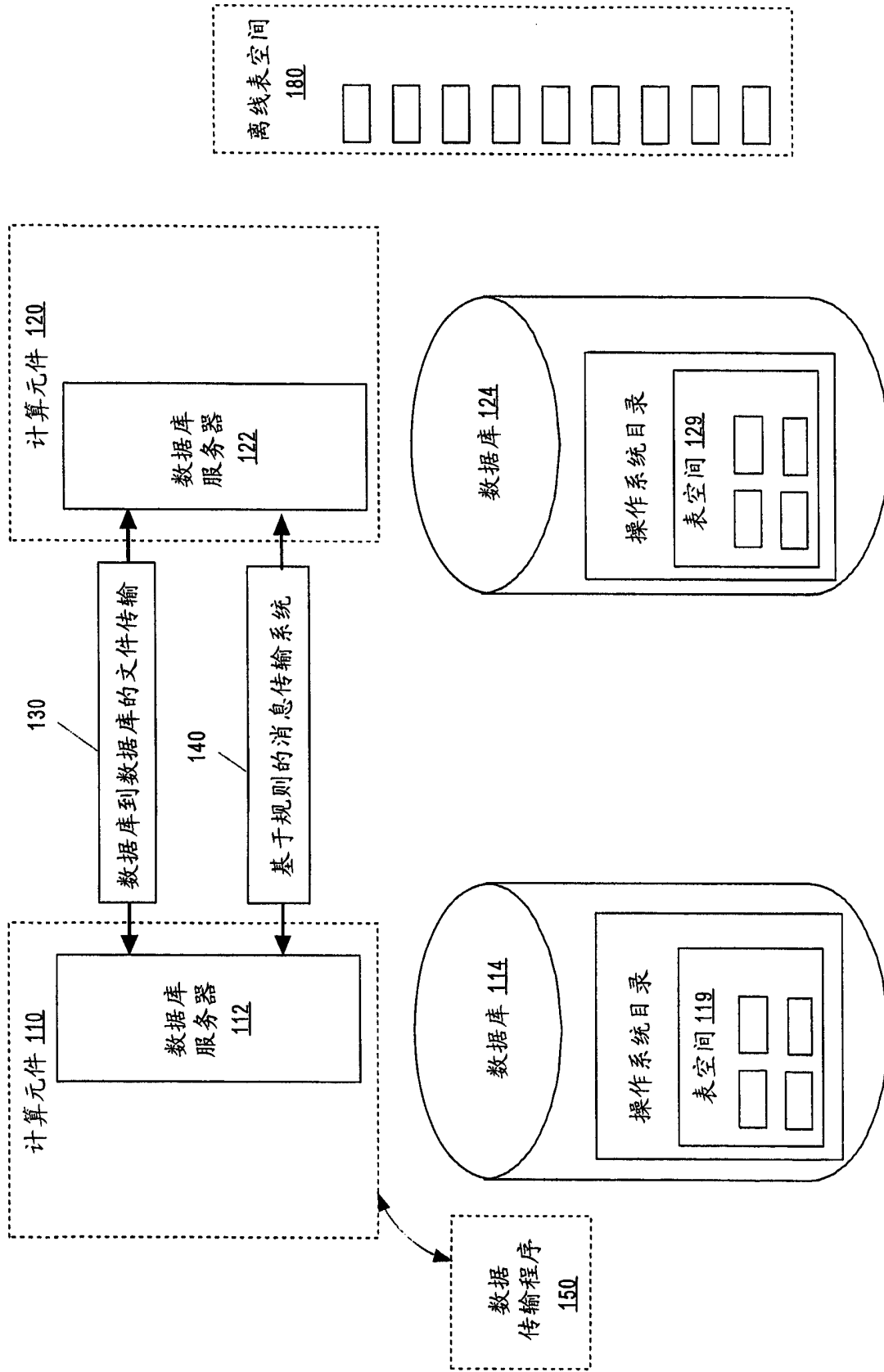
局域网 **622** 相连。例如，通信接口 **618** 可以是综合业务数字网（Integrated Services Digital Network, ISDN）卡或者调制解调器，用于提供到相应类型电话线的数据通信连接。又如，通信接口 **618** 可以是局域网（Local Area Network, LAN）卡，用于提供至兼容 LAN 的数据通信连接。也可以使用无线链路。在任何这样的实施方式中，通信接口 **618** 均发送和接收承载表示各种信息的数字数据流的电信号、电磁信号和光学信号。

网络链路 **620** 通常可通过一个或者多个网络提供数据通信给其它数据装置。例如，网络链路 **620** 可通过局域网 **622** 提供与主机 **624** 连接，或者提供与互联网服务提供商（Internet Service Provider, ISP）**626** 操作的数据设备的连接。ISP **626** 进而通过目前通称为“互联网”**628** 的全球分组数据通信网络提供数据通信服务。局域网 **622** 和互联网 **628** 都使用承载数字数据流的电信号、电磁信号或光学信号。通过各种网络的信号、和网络链路 **620** 上的信号、和通过通信接口 **618** 的信号，是传输信息的载波的典型形式，这些信号都传送数字数据给计算机系统 **600** 或者传送来自计算机系统 **600** 的数字数据。

计算机系统 **600** 能通过网络、网络链路 **620**、和通信接口 **618** 发送消息和接收数据（包括程序代码）。在互联网实例中，服务器 **630** 可通过互联网 **628**、ISP **626**、局域网 **622**、和通信接口 **618** 传送关于应用程序的所请求代码。

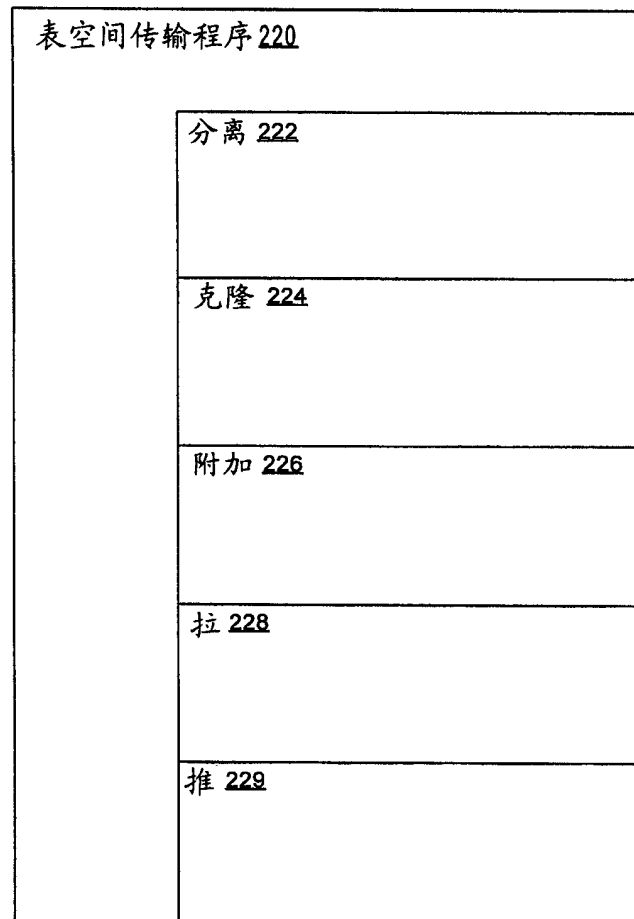
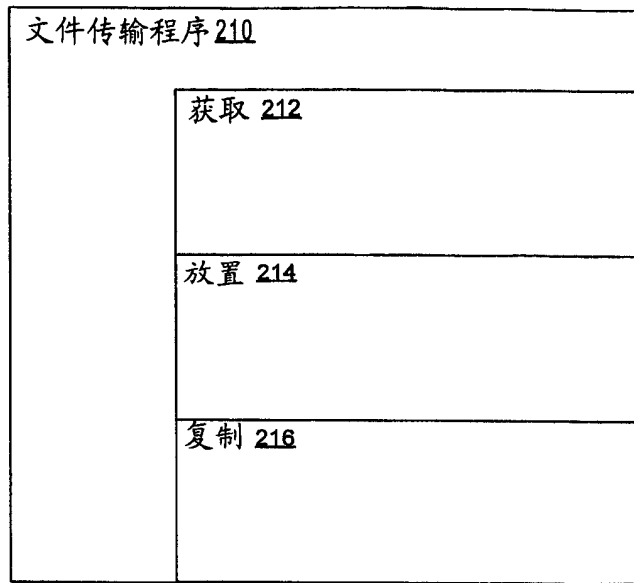
当代码被接收和/或储存在存储装置 **610** 上或者其它非易失性存储器上用于随后执行时，处理器 **604** 可执行所接收到的代码。以此方式，计算机系统 **600** 可以获得载波形式的应用程序代码。

以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。



分布式数据库系统 101

图 1



数据传输程序 150

图 2

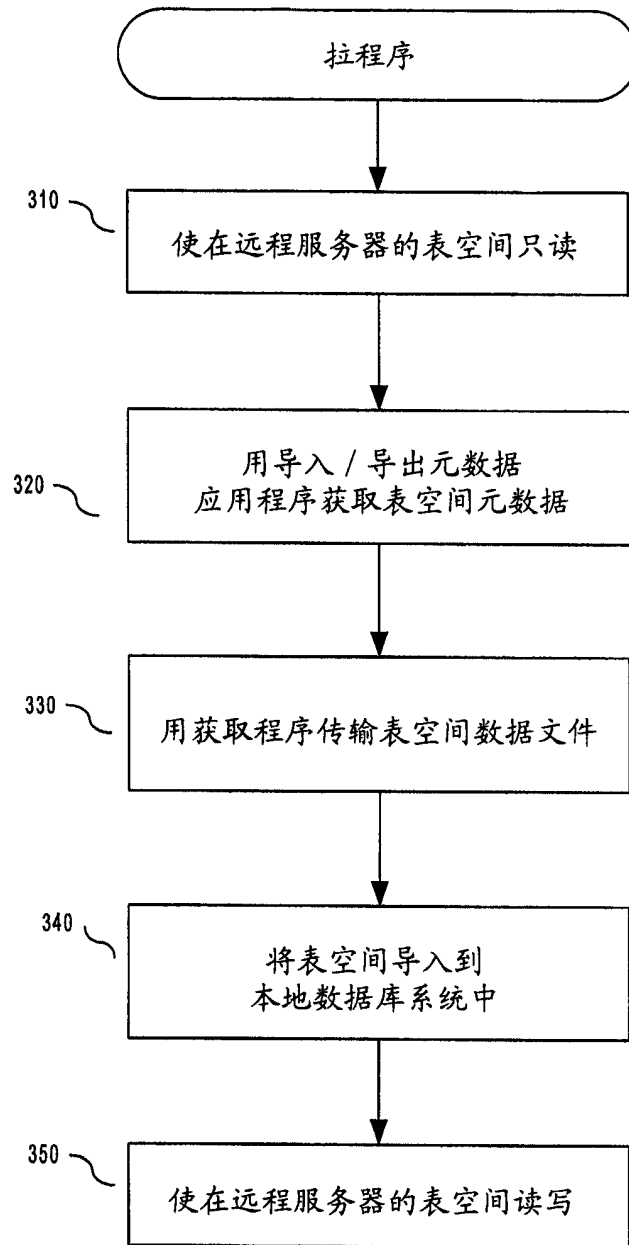
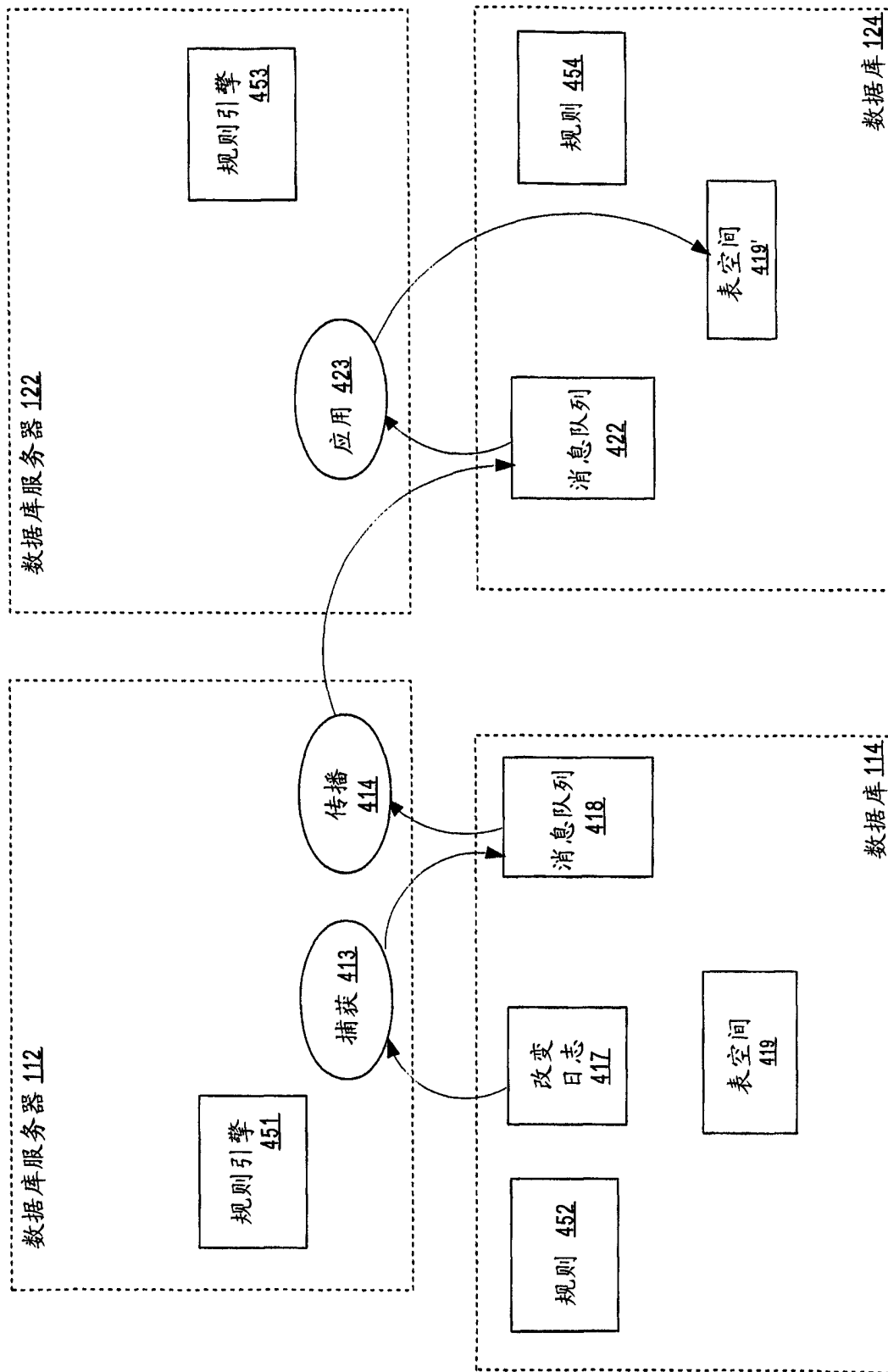


图 3



分布式数据库系统 101

图 4

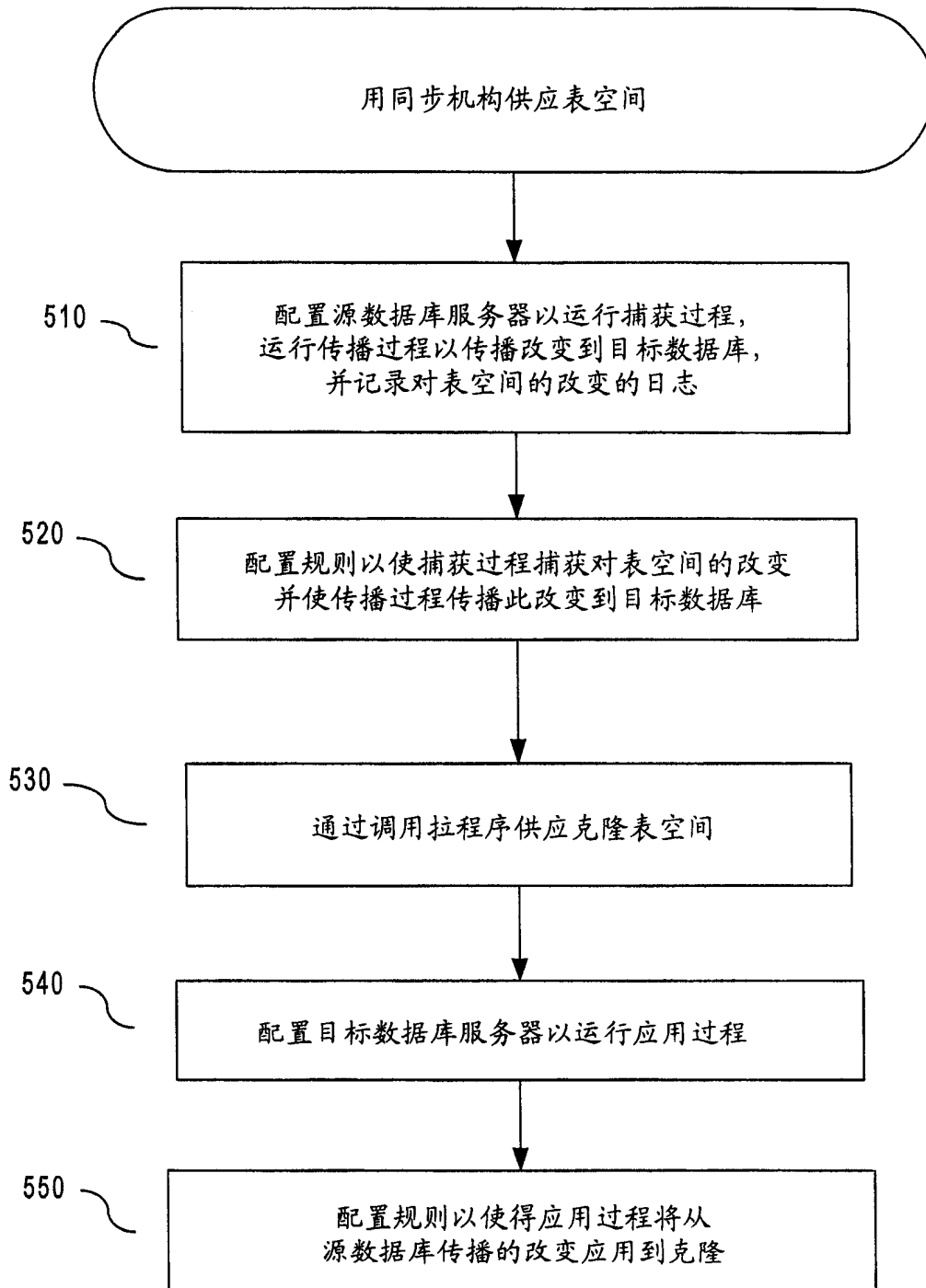


图 5

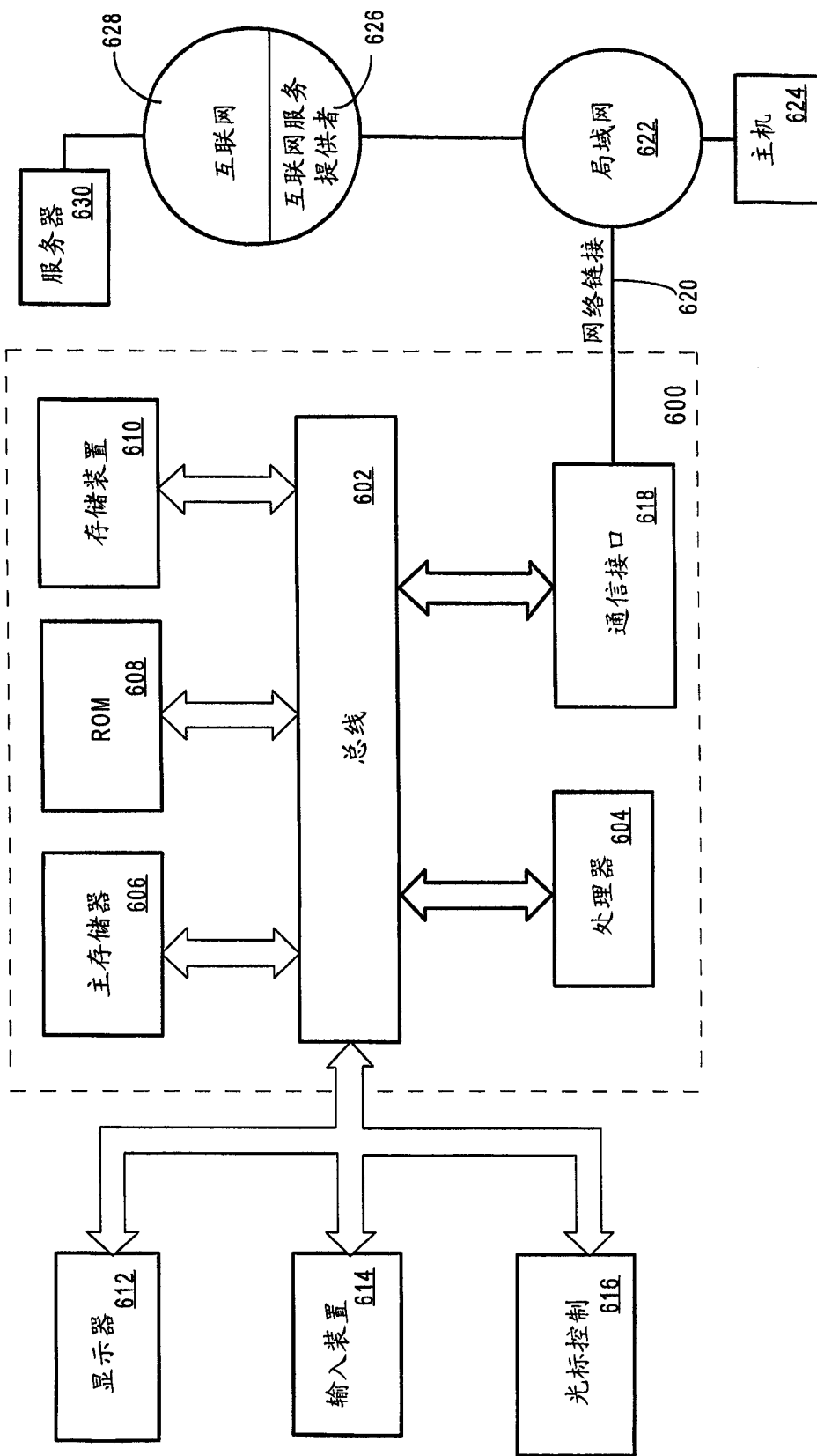


图6