



(12) 发明专利

(10) 授权公告号 CN 101390075 B

(45) 授权公告日 2015. 09. 02

(21) 申请号 200780006331. 4

(22) 申请日 2007. 02. 13

(30) 优先权数据

11/359, 276 2006. 02. 22 US

(85) PCT国际申请进入国家阶段日

2008. 08. 22

(86) PCT国际申请的申请数据

PCT/US2007/004048 2007. 02. 13

(87) PCT国际申请的公布数据

W02007/100509 EN 2007. 09. 07

(73) 专利权人 微软技术许可有限责任公司

地址 美国华盛顿州

(72) 发明人 J·李

(74) 专利代理机构 上海专利商标事务所有限公

司 31100

代理人 顾嘉运

(51) Int. Cl.

G06F 15/16(2006. 01)

G06F 17/00(2006. 01)

G06F 17/40(2006. 01)

(56) 对比文件

US 2004/0162808 A1, 2004. 08. 19, 说明书第 [0029]–[0167] 段.

US 2003/0145093 A1, 2003. 07. 31, 说明书 [0039]–[0075] 段.

审查员 王宁

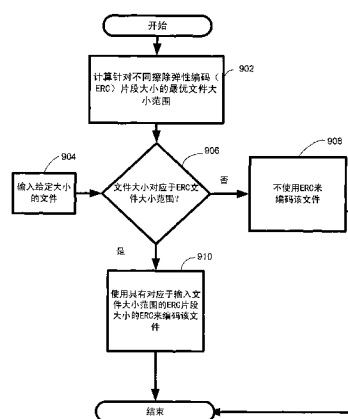
权利要求书3页 说明书11页 附图10页

(54) 发明名称

可靠、高效的对等存储

(57) 摘要

一种使用根据分发的文件的大小改变用于编码的片段的数量的自适应擦除弹性编码 (ERC) 的自适应编码存储系统。自适应 ERC 可极大地提高 P2P 存储的效率和可靠性。还可实现 P2P 存储应用的多个过程。在一个实施例中, 小的、动态数据文件被转向更可靠的对等体或甚至服务器, 而大的及静态文件利用不可靠对等体的存储容量来存储。同样地, 为了平衡的贡献和收益, 对等体应当主存与其存储在 P2P 网络中相同数量的内容。结果, 不可靠的对等体被允许分发较少的数据, 而更可靠的对等体被允许分发较多的数据。同样地, 较小的文件被分配较高的分发成本, 而较大的文件被分配较低的分发成本。



1. 一种用于编码文件以便存储在对等网络中的计算机实现的方法, 包括以下过程动作:

计算对应于不同擦除弹性编码 ERC 片段数的最优文件大小范围, 其中每个片段数都是对于相应的文件大小范围的最优片段数 (902);

输入给定文件大小的文件 (904);

如果所述文件大小小于对应于最小 ERC 片段数为 2 的文件大小范围, 则不使用擦除弹性编码来编码所述文件 (906、908);

如果所述输入文件的文件大小对应于一文件大小范围, 则使用擦除弹性编码和对应于所述输入文件的文件大小范围的最优片段数来编码所述文件。

2. 如权利要求 1 所述的计算机实现的方法, 其特征在于, 还包括将所述编码的文件发送到对等网络中的一个或多个对等体的过程动作。

3. 如权利要求 1 所述的计算机实现的方法, 其特征在于, 还包括根据对等体可靠性和期望的文件内容可靠性来计算将向其存储所述编码的文件的对等体的数量。

4. 如权利要求 1 所述的计算机实现的方法, 其特征在于, 计算对应于不同擦除弹性编码 ERC 片段数的最优文件大小范围包括以下过程动作:

按照适于特定 ERC 片段数的最优文件大小确定不同片段数之间的边界。

5. 如权利要求 1 所述的计算机实现的方法, 其特征在于, 还包括以下过程动作:

获得等于或大于所述文件被拆分成以进行编码的片段数的所述文件的一组编码的文件片段; 以及

如果所述文件是擦除弹性编码的, 则用擦除弹性解码来解码所述编码的文件片段以获得所述编码的文件的解码版本; 以及

如果所述文件不是擦除弹性编码的, 则不用擦除弹性解码来解码所述编码的片段以获得所述编码的文件的解码版本。

6. 如权利要求 1 所述的计算机实现的方法, 其特征在于, 在编码所述文件时所使用的擦除弹性编码是 Reed Solomon 编码。

7. 如权利要求 6 所述的计算机实现的方法, 其特征在于:

如果所述文件大小小于约 10KB, 则不使用擦除弹性编码;

如果所述文件大小约是 10KB 到 33KB, 则所述最优片段数是 2;

如果所述文件大小约是 33KB 到 100KB, 则所述最优片段数是 4;

如果所述文件大小约是 100KB 到 310KB, 则所述最优片段数是 8;

如果所述文件大小约是 310KB 到 950KB, 则所述最优片段数是 16;

如果所述文件大小约是 950KB 到 2.9MB, 则所述最优片段数是 32;

如果所述文件大小约是 2.9MB 到 8.9MB, 则所述最优片段数是 64;

如果所述文件大小约是 8.9MB 到 26MB, 则所述最优片段数是 128; 以及

如果所述文件大小大于约 26MB, 则所述最优片段数是 256。

8. 一种用于编码文件以便存储在对等网络中的系统, 包括:

用于计算对应于不同擦除弹性编码 ERC 片段数的最优文件大小范围的装置, 其中每个片段数都是对于相应的文件大小范围的最优片段数;

用于输入给定文件大小的文件的装置;

用于如果所述文件大小小于对应于最小 ERC 片段数为 2 的文件大小范围,则不使用擦除弹性编码来编码所述文件的装置;

用于如果所述输入文件的文件大小对应于一文件大小范围,则使用擦除弹性编码和对应于所述输入文件的文件大小范围的最优片段数来编码所述文件的装置。

9. 一种用于提高对等网络的存储可靠性和效率的方法,包括:

确定用于用擦除弹性编码来编码给定大小的文件的最优片段数 (1010);

如果所述用于用擦除弹性编码来编码文件的最优片段数是 1,则不用擦除弹性编码来编码所述文件 (1010);以及

如果所述最优片段数大于 1,则通过将所述文件分成最优数量的片段并用擦除弹性编码编码所述文件 (1010) 来编码所述文件。

10. 如权利要求 9 所述的方法,其特征在于,还包括根据对等体可靠性和期望的文件内容可靠性来计算将向其存储所述编码的文件的对等体的数量。

11. 如权利要求 10 所述的方法,其特征在于,还包括将用擦除弹性编码编码的文件分发到网络上的一个或多个对等体。

12. 如权利要求 11 所述的方法,其特征在于,所述将用擦除弹性编码编码的文件分发到网络上的一个或多个对等体还包括:

确定所述对等网络中的每个对等体的可靠性;

确定所述文件的大小;

以及使用按照文件大小所确定的具有适当可靠性的一个或多个对等体来分发所述文件。

13. 如权利要求 12 所述的方法,其特征在于,所述将用擦除弹性编码编码的文件分发到网络上的一个或多个对等体还包括:

如果所述文件是大文件,则使用其可靠性低于给定阈值的对等体来分发该大文件;以及

如果所述文件不是大文件,则使用其可靠性高于给定阈值的对等体来分发该文件。

14. 如权利要求 11 所述的方法,其特征在于,所述将用擦除弹性编码编码的文件分发到网络上的一个或多个对等体还包括:

确定所述对等网络中的每个对等体的可靠性;

确定所述文件是否是静态的;

如果所述文件是静态的,则使用其可靠性低于给定阈值的对等体来分发该文件;以及

如果所述文件不是静态的,则使用其可靠性高于给定阈值的对等体来分发该文件。

15. 如权利要求 11 所述的方法,其特征在于,所述将用擦除弹性编码编码的文件分发到网络上的一个或多个对等体还包括:

确定所述对等网络中的每个对等体的可靠性;

监视所述文件的改变;

首先将所述文件分发到较可靠的对等体;

如果所述文件被观察到没有改变,则将所述文件重新分发到较不可靠的对等体。

16. 如权利要求 9 所述的方法,其特征在于,还包括使用服务器来提高所述对等网络的效率,包括:

将所述对等网络中的所有动态文件备份到所述服务器；

周期性地令所述对等网络中的对等体和服务器检查备份到所述服务器的动态文件是否已经改变；

如果所述动态文件还没有改变，则将这些文件指定为静态的并且将它们一起捆绑成一大文件；以及

用擦除弹性编码将所述大文件分发到所述对等网络。

17. 如权利要求 9 所述的方法，其特征在于，所述确定用于用擦除弹性编码来编码给定大小的文件的最优片段数包括：

确定对应于每个可能的擦除弹性编码片段数的最优文件大小范围，其中每个片段数是对应于相应范围的最优片段数；

确定输入文件的大小落入哪一文件大小范围；以及

使用对应于所述输入文件的大小所落入的最优文件大小范围的片段数作为所述最优片段数。

18. 一种用于解码存储在所述对等网络中的编码的文件的计算机实现的方法，包括以下过程动作：

检索等于或大于用于编码所述文件的片段的数量的该编码的文件的一组片段，其中所述文件是用对应于给定文件大小的最优片段数来擦除弹性编码的；以及

用擦除弹性解码来解码所述编码的片段以获得所述编码的文件的解码版本。

19. 如权利要求 18 所述的计算机实现的方法，其特征在于，所述编码的片段中的至少某一些是从所述对等网络中的一个或多个对等体的存储介质中检索的。

20. 如权利要求 18 所述的计算机实现的方法，其特征在于，所述编码的片段中的至少某一些是从所述对等网络上的服务器的存储介质中检索的。

可靠、高效的对等存储

[0001] 背景

[0002] 在对等 (P2P) 应用中,当对等体加入 P2P 服务时携带网络带宽和 / 或硬盘驱动器存储资源。随着对 P2P 系统的需求增长,系统的容量也随之增长。这与客户机 - 服务器系统完全相反,在客户机 - 服务器系统中,服务器的容量是固定的,并且由客户机 - 服务器系统的提供方来支付。结果,运行 P2P 系统比运行客户机 - 服务器系统更经济,并且因为 P2P 系统是可伸缩的,所以它是优越的。

[0003] 在 P2P 系统中,对等体不仅贡献带宽,还对服务其他对等体的存储空间做出贡献。由对等体贡献的整体存储空间形成了分布式存储云。数据可被存储到云中,并且可从中检索。P2P 存储可用于大量的应用。分布式备份就是其中之一。对等体可将其自己的数据备份到 P2P 云中。当对等体发生故障时,数据可从云中还原。另一个 P2P 应用是分布式数据访问。因为客户机可从多个保持数据的对等体中同时检索数据,所以,P2P 检索能够具有相比于从单个源中检索数据更高的吞吐量。另一个应用是点播电影观看。媒体服务器可抢先在 P2P 云中散播电影文件。当客户机正在观看该电影时,它可从 P2P 云和服务器处流传送该电影,因此减少服务器负载,减少在网络主干上的通信量并提高流电影的质量。

[0004] 尽管在 P2P 网络中的对等体可像服务器那样运作,但是它们在一个重要的方面与商用 web/ 数据库服务器不同:可靠性。因为对等体通常是用其多余的硬盘驱动器空间和空闲带宽来支持 P2P 应用的普通计算机,所以它远没有典型的服务器可靠。用户有时可选择关闭对等计算机或 P2P 应用程序。例如,大文件上传 / 下载的必要需求可使对等体极度缺乏用于 P2P 活动的必要带宽。对等计算机由于需要升级或修补软件 / 硬件,或由于病毒攻击可能是脱机的。对等体的计算机硬件和网络链接也固有地远没有为可靠性而设计的典型的服务器计算机和其商用网络链接可靠。然而,商用服务器 / 服务器群集是为“六个 9”的可靠性 (具有 10^{-6} 的故障率,以该故障率,每年允许大约 30 秒的停机时间) 而设计的,优良的消费者对等体可能只具有“二个九”的可靠性 (10^{-2} 的故障率或每天大约 15 分钟的停机时间),并且对等体只具有 50% (时间下降一半) 或甚至 10% (时间下降 90%) 的可靠性并非不常见。

[0005] 例如 P2P 备份和数据检索等大多数 P2P 应用想要维持与服务器的等级 (“六个 9”的可靠性) 相同的 P2P 存储可靠性等级。该挑战在于如何使用对等体的最少的带宽和存储资源来构建可靠、高效的 P2P 存储。

[0006] 概述

[0007] 提出了一种用于在对等 (P2P) 网络中高效且可靠地存储数据的自适应编码存储系统和方法。该自适应编码存储系统和方法基于存储和分发的文件大小来调整擦除弹性编码 (ERC) 的多个片段、ERC 片段数。

[0008] 该自适应编码存储系统的多个实施例利用各过程来提高 P2P 网络的效率和可靠性。例如,在一个实施例中,如果服务器组件支持是可用的,则将较小的、动态的数据转向更可靠的对等体或甚至服务器。同样地,在另一个实施例中,对于平衡的 P2P 网络,不可靠并且正在分发较小文件的对等体被允许分发较少的数据。

[0009] 注意,虽然背景一节中所描述的现有的对等存储和分发系统中的上述限制可由根据本发明的自适应编码存储系统的一个具体实现来解决,但是该系统和过程决不限于只解决所述缺点中的任一个或所有的实现。相反,如将从以下描述中变得明显的,本系统和过程具有广泛得多的应用。

[0010] 应当注意,提供本概述以便使用简化形式介绍下面在详细描述中进一步描述的一些概念。本概述不旨在标识所要求保护的的主题的关键特征或必要特征,也不旨在用于帮助确定所要求保护的的主题的范围。

附图说明

[0011] 参考以下描述、所附权利要求书以及附图,将更好地理解自适应编码存储系统的具体特征、方面、以及优点,附图中:

[0012] 图 1 是描绘构成实现此处所描述的自适应编码存储系统和方法的示例性系统的通用计算设备的概括系统图。

[0013] 图 2 示出了可用于此处所描述的自适应编码存储系统和方法的示例性对等 (P2P) 网络。

[0014] 图 3 提供了示出实现期望的 10^{-6} 的可靠性的信息存储对等体的数量的曲线图。

[0015] 图 4 提供了示出对等体可靠性以及期望的重复率的曲线图。

[0016] 图 5 提供了示出使用擦除弹性编码来实现期望的 10^{-6} 的可靠性所必需的信息存储对等体的数量的曲线图。

[0017] 图 6 提供了 ERC 片段数以及相关联的适于 P2P 网络中的信息存储的文件大小的曲线图。

[0018] 图 7 提供了描绘在具有自适应 ERC 和固定 ERC 的 P2P 配置中的对等体之间的带宽使用的曲线图 (在对等体可靠性 = 50% 的情况下)。

[0019] 图 8 提供了描绘在具有自适应 ERC 和固定 ERC 的 P2P 配置中的对等体之间的带宽使用的曲线图 (在对等体可靠性 = 99% 的情况下)。

[0020] 图 9 描绘了自适应编码存储过程的一个实施例。

[0021] 图 10 描绘了示出在 P2P 网络中如何利用自适应编码存储技术的示例性操作流程

图。

[0022] 图 11 描绘了实现优化 P2P 网络的存储效率的过程的自适应编码存储系统和过程的实施例。

[0023] 图 12 描绘了实现优化 P2P 系统的存储效率的过程的自适应编码存储系统和方法的另一个实施例。

[0024] 图 13 描绘了利用具有服务器支持的 P2P 备份的自适应编码存储系统和方法的实施例。

[0025] 详细描述

[0026] 在本自适应编码存储系统的优选实施例的以下描述中,对附图进行了参考,附图构成了实施例的一部分且在其中作为说明示出了可在其中实践自适应编码存储系统的具体实施例。可以理解,可以使用其它实施例并且可以做出结构上的改变而不背离本自适应编码存储系统的范围。

[0027] 1.0 示例性操作环境

[0028] 图 1 示出了可在其上实现本发明的合适的计算系统环境 100 的示例。计算系统环境 100 仅为合适的计算环境的一个示例,并非对本发明的使用范围或功能提出任何局限。也不应该把计算环境 100 解释为对示例性操作环境 100 中示出的任一组件或其组合有任何依赖性要求。

[0029] 本发明可用各种其它通用或专用计算系统环境或配置来操作。适用于本发明的公知的计算系统、环境和 / 或配置的示例包括,但不限于,个人计算机、服务器计算机、诸如蜂窝电话和 PDA 等手持式、膝上型或移动计算机或通信设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费电子产品、网络 PC、小型机、大型计算机、包含上述系统或设备中的任一个的分布式计算环境等。

[0030] 本发明可在诸如程序模块等由计算机执行的计算机可执行指令结合包括话筒阵列 198 的组件在内的硬件模块的一般上下文中描述。一般而言,程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等。本发明也可以在其中任务由通过通信网络链接的远程处理设备执行的分布式计算环境中实现。在分布式计算环境中,程序模块可以位于包括存储器存储设备在内的本地和远程计算机存储介质中。参考图 1,用于实现本发明的一个示例性系统包括计算机 110 形式的通用计算设备。

[0031] 计算机 110 的组件可以包括,但不限于:处理单元 120、系统存储器 130 和将包括系统存储器在内的各种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、外围总线以及使用各类总线体系结构中的任一种的局部总线。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、微通道体系结构 (MCA) 总线、增强型 ISA (EISA) 总线、视频电子技术标准协会 (VESA) 局部总线、以及外围部件互连 (PCI) 总线 (也称为小背板 (Mezzanine) 总线)。

[0032] 计算机 110 通常包括各种计算机可读介质。计算机可读介质可以是可由计算机 110 访问的任何可用介质,且包括易失性和非易失性介质、可移动和不可移动介质。作为示例,而非限制,计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其它数据这样的信息的任意方法或技术来实现的易失性和非易失性、可移动和不可移动介质。

[0033] 计算机存储介质包括,但不限于, RAM、ROM、PROM、EPROM、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; 磁带盒、磁带、磁盘存储或其它磁性存储设备; 或能用于存储所需信息且可以由计算机 100 访问的任何其它介质。通信介质通常以诸如载波或其它传输机制等已调制数据信号来体现计算机可读指令、数据结构、程序模块或其它数据,且包含任何信息传递介质。术语“已调制数据信号”指的是这样一种信号,其一个或多个特征以在信号中编码信息的方式被设定或更改。作为示例而非限制,通信介质包括有线介质,诸如有线网络或直接线连接,以及无线介质,诸如声学、RF、红外线和其它无线介质。上述的任意组合应该包含在计算机可读介质的范围内。

[0034] 系统存储器 130 包括易失性和 / 或非易失性存储器形式的计算机存储介质,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) 包含有助于诸如启动时在计算机 110 中元件之间传递信息的基本例程,它通常被存储在 ROM 131

中。RAM 132 通常包含处理单元 120 可以立即访问和 / 或目前正在操作的数据和 / 或程序模块。作为示例而非局限,图 1 示出了操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137。

[0035] 计算机 110 也可以包括其它可移动 / 不可移动、易失性 / 非易失性计算机存储介质。仅作为示例,图 1 示出了从不可移动、非易失性磁介质中读取或向其写入的硬盘驱动器 141,从可移动、非易失性磁盘 152 中读取或向其写入的磁盘驱动器 151,以及从诸如 CD ROM 或其它光学介质等可移动、非易失性光盘 156 中读取或向其写入的光盘驱动器 155。可以在示例性操作环境中使用的其它可移动 / 不可移动、易失性 / 非易失性计算机存储介质包括,但不限于,盒式磁带、闪存卡、数字多功能盘、数字录像带、固态 RAM、固态 ROM 等。硬盘驱动器 141 通常由不可移动存储器接口,诸如接口 140 连接至系统总线 121,磁盘驱动器 151 和光盘驱动器 155 通常由可移动存储器接口,诸如接口 150 连接至系统总线 121。

[0036] 上文讨论并在图 1 中示出的驱动器及其关联的计算机存储介质为计算机 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。例如,在图 1 中,硬盘驱动器 141 被示为存储操作系统 144、应用程序 145、其它程序模块 146 和程序数据 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 和程序数据 137 相同或不同。操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147 在这里被标注了不同的标号是为了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 和定点设备 161 (通常指鼠标、跟踪球或触摸板) 向计算机 110 输入命令和信息。

[0037] 其它输入设备 (未示出) 可包括操纵杆、游戏手柄、圆盘式卫星天线、扫描仪、无线电接收机、以及电视或广播视频接收机等等。这些和其它输入设备通常通过耦合至系统总线 121 的有线或无线用户输入接口 160 连接至处理单元 120,但是也可以通过其它常规的接口和总线结构来连接,如并行端口、游戏端口、通用串行总线 (USB)、IEEE 1394 接口、蓝牙 TM 无线接口、IEEE 802.11 无线接口等等。此外,计算机 110 还可包括语音或音频输入设备,诸如话筒或话筒阵列 198、以及经由音频接口 199 连接的扬声器 197 或其它声音输出设备,音频接口同样包括常规的有线或无线接口,诸如并行、串行、USB、IEEE 1394、蓝牙 TM 等等。

[0038] 监视器 191 或其它类型的显示设备也经由诸如视频接口 190 等接口连接到系统总线 121。除监视器以外,计算机也可以包括其它外围输出设备,诸如打印机 196,它可以通过输出外围接口 195 连接。

[0039] 计算机 110 可使用至一个或多个远程计算机,诸如远程计算机 180 的逻辑连接在网络化环境中操作。远程计算机 180 可以是个人计算机、服务器、路由器、网络 PC、对等设备或其它常见的网络节点,且通常包括上文相对于计算机 110 描述的许多或所有元件,尽管在图 1 中只示出存储器存储设备 181。图 1 中所示的逻辑连接包括局域网 (LAN) 171 和广域网 (WAN) 173,但也可以包括其它网络。这样的联网环境在办公室、企业范围计算机网络、内联网和因特网中是常见的。

[0040] 当在 LAN 联网环境中使用时,计算机 110 通过网络接口或适配器 170 连接至 LAN 171。当在 WAN 联网环境中使用时,计算机 110 通常包括调制解调器 172 或用于通过诸如因特网等 WAN 173 建立通信的其它装置。调制解调器 172 可以是内置或外置的,它可以通过用户输入接口 160 或其它合适的机制连接至系统总线 121。在网络化环境中,相对于计算机

110 所描述的模块或其部分可以存储在远程存储器存储设备中。作为示例而非局限，图 1 示出远程应用程序 185 驻留在存储器设备 181 上。可以理解，示出的网络连接是示例性的，并且可以使用在计算机之间建立通信链路的其它手段。

[0041] 一般而言，自适应编码存储系统在诸如图 2 所示的网络等的 P2P 网络中操作。对于特定的数据流会话，“服务器”200 被定义为在 P2P 网络中最初发起数据或流媒体的节点；“客户机”（或接收者）210 被定义为当前请求数据的节点；并且“服务对等体 220”被定义为向客户机供应数据的完整或部分副本的节点。

[0042] 一般而言，服务器 200、客户机 210 以及服务对等体 220 都是连接到诸如因特网等的网络的最终用户节点。因为服务器 200 始终能够供应数据，所以服务器节点还担当服务对等体 220。服务器节点 200 还能够执行不能由服务对等体 220 执行的管理功能，例如，维护可用的服务对等体列表、执行数字权限管理（DRM）功能等等。此外，如同传统的 P2P 方案一样，由于部署了越来越多的对等体节点 220，因此此处所描述的自适应编码存储系统从提高效率中获益。具体地，随着对等体节点 220 的数量的增长，数据服务器 200 上的负载将随之减少，由此变得以较少的代价运行，而在特定数据传输会话期间每个客户机节点 210 将能够接收好得多的数据质量。

[0043] 此外，应当清楚，特定节点的角色可能改变。例如，一个特定节点可在一个特定数据传输中充当客户机 210 的角色，而在另一个会话中充当服务对等体 220 的角色。此外，特定节点能够同时充当客户机节点 210 和服务器 200 或服务对等体 220 两者的角色以同时发送一个或多个数据文件，或这些文件的一部分，并从一个或多个其他服务对等体接收其他数据。

[0044] 在数据传输期间，客户机 200 首先定位保持所需数据中的某一些或全部的多个邻近的对等体 220，并且然后从（可能包括服务器 200 的）多个对等体接收数据。结果，每个服务对等体 220 通过服务客户机 210 的一部分下载请求来减轻总上传负担而协助服务器 200。结果，由于当有许多服务对等体 220 协助服务器 200 时有相当高的服务带宽可用，所以特别是在存在许多客户机的情况下，客户机 210 经常能够接收好得多的数据质量。

[0045] 现在已讨论了示例性操作环境，本描述一节的其余部分将致力于描述实现自适应编码存储系统和过程的程序模块。

[0046] 2.0 可靠、高效的对等存储

[0047] 自适应编码存储系统提供自适应地确定是否使用 ERC 编码并且利用将用于给定文件大小的 ERC 编码的最优片段数以得到最优可靠性和效率的自适应擦除弹性编码（ERC）方案。用于文件的 ERC 编码的片段数将出于本讨论的目的被称为“ERC 片段数”。以下各段提供对对等（P2P）存储效率和可靠性以及在 P2P 网络中对 ERC 的使用的讨论，以及对所使用的 ERC 片段数的讨论。然后讨论自适应编码存储系统和过程的各实施例。

[0048] 2.1 P2P 存储的可靠性：数据冗余

[0049] 为具有不可靠部分的系统带来可靠性的特定解决方案是使用冗余。如果网络上的每个单独的对等体具有可靠性 p ，则要实现期望的可靠性 p_0 ，一个对等体可简单地将信息复制到 n 个对等体：

[0050]
$$n = \log(1-p_0) / \log(1-p), (1)$$

[0051] 其中 n 是保存信息的对等体的数量。在检索时间，客户机可挨个联系信息存储对

等体。只要信息存储对等体中的一个上线,就能够可靠地检索信息。

[0052] 尽管实现了可靠性,但是简单的复制策略并不是高效的。图 3 绘出了实现“六个 9”可靠性所需的信息存储对等体的数量的曲线。对于 50% 的对等体可靠性,一个对等体需要将信息复制并存储到 20 个对等体。这导致需要 20 倍的带宽和存储空间来分发和存储这些信息。显然,牺牲了效率以交换信息可靠性。

[0053] 2. 2P2P 的擦除弹性编码

[0054] 为了在仍维持相同的可靠性的同时提高效率,ERC 可以是有用的工具。ERC 将原始文件分成 k 个原始片段 $\{x_i\}$, $i = 0, \dots, k-1$, 每个片段都是 Galois 域 $GF(q)$ 上的一个向量,其中 q 是该域的阶数。例如正在编码 64KB 长的文件,如果使用 $q = 216$ 且 $k = 16$,则每个片段将会是 4KB,且将由 2K 的字组成,且每个字都是 $GF(216)$ 的一个元素。ERC 然后从原始片段生成编码的片段。经 ERC 编码的片段是由以下运算形成的:

$$[0055] \quad c_j = G_i \square [x_0 \ x_1 \cdots x_{k-1}]^t, \quad (2)$$

[0056] 其中 c_j 是编码的片段, G_i 是 k 维生成向量,并且等式 (2) 是全部都在 $GF(q)$ 上的矩阵乘法。在解码时间,对等体收集 m 个编码的片段,其中 m 是等于或略大于 k 的数字并且尝试解码 k 个原始片段。这等价于求解等式:

$$[0057] \quad \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{m-1} \end{bmatrix} = \begin{bmatrix} G_0 \\ G_1 \\ \vdots \\ G_{k-1} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{k-1} \end{bmatrix}, \quad (3)$$

[0058] 如果由生成向量形成的矩阵具有满秩 k ,则可恢复原始消息。

[0059] 有许多可用的 ERC。特别感兴趣的一种是 Reed-Solomon (RS) 码。RS 码使用结构化的生成向量,并且是最大可分距离 (MDS)。结果,任何 k 个不同的编码的片段将能够解码原始片段。RS 码的另一个优点是可通过生成向量的索引 i 来容易地标识和管理编码的片段,因此便于检测重复的 RS 码。在以下 ERC 的讨论中,假设使用 RS 码。然而,该自适应编码存储系统可用任何数量的传统 ERC 来实现。

[0060] 2. 3ERC: 片段数量

[0061] 通过在 P2P 存储中使用 ERC,数据文件被分发给更多的对等体,但是每个对等体只需要存储是原始文件的 $1/k$ 大小的一个编码的片段,导致总体减少了实现相同等级的可靠性所需的带宽和存储空间,并因此提高了效率。令 n_1 为需要向其分发编码的片段以实现某一期望的可靠性等级的对等体数量。因为 RS 码是 MDS 码,所以保持 k 个不同的编码的片段的 k 个对等体将足以恢复原始文件。恰好有 m 个对等体可用的概率可用过二项式分布来计算:

$$[0062] \quad p(m, n_1) = \binom{n_1}{m} p^m (1-p)^{n_1-m}. \quad (4)$$

[0063] 因此可如下从 p 、 p_0 和 k 计算 n_1 :

$$[0064] \quad n_1 = \arg \min_o \left\{ \sum_{m < k} \binom{o}{m} p^m (1-p)^{o-m} < 1 - p_0 \right\}. \quad (5)$$

[0065] 重复率 r 被定义为:

$$[0066] \quad r = n_1/k \quad (6)$$

[0067] 重复率 r 是良好的效率指示符, 因为文件的 r 个副本需要被分发和存储到 P2P 云中。

[0068] 图 4 示出了对于不同 ERC 片段数 k 实现“六个 9”可靠性所期望的重复率。可观察到对 ERC 的使用极大地降低了所需的重复率。比较非 ERC ($k = 1$) 和 ERC 片段数 $k = 256$, 对于 10% 的对等体可靠性所需的重复率从 $r = 132$ 减少到 $r = 13.1$, 对于 50% 的对等体可靠性所需的重复率从 $r = 20$ 减少到 $r = 2.5$, 而对于 99% 的对等体可靠性所需的重复率从 $r = 3$ 减少到 $r = 1.05$ 。ERC 可提高效率而不会牺牲可靠性。

[0069] 还可观察到更大的 ERC 片段数进一步降低了重复率。对于 50% 的对等体可靠性, 从 $k = 8$ 到 16、32、64、128 和 256 导致重复率从 $r = 5.75$ 减少到 4.375、3.53、3.02、2.68 和 2.48。相应的效率提高分别是 24%、19%、15%、11% 和 8%。这似乎建议应当使用大的 ERC 片段数来得到更到的效率。

[0070] 然而, 更大的 ERC 片段数意味着需要更多的对等体来存储和检索编码的片段。如图 5 所示, 绘出了需要保持编码的片段以实现“六个 9”的可靠性的对等体数量的曲线。再次对于 50% 对等体可靠性, 从 $k = 8$ 到 16、32、64、128 和 256 将信息存储对等体的数量从 $n_1 = 46$ 增加到 70、113、193、343 和 630。每次 k 的翻倍导致存储信息所需的对等体多出 52%、61%、71%、78%、84%。 k 的翻倍也需要在信息检索期间将联系的对等体数量至少翻倍。

[0071] 在大多数实际的 P2P 网络中, 在对等体之间建立连接需要不平凡数量的开销。开销的一部分归结于检索正确的对等体身份并找到正确的路由路径 (例如, 通过分布式散列表 (DHT))。开销的另一部分是由于在对等体之一或两者位于 NAT 之后的情况下调用某些网络地址翻译 (NAT) 算法的需求, 例如, STUN (经 NAT 的简单 UDP 穿越)。假设在对等体之间建立连接的平均开销是 $overhead$ (在本示例中设置成 16KB), 则可计算存储 s 大小的文件所需的总网络带宽为:

[0072] $store_bandwidth = s * r + n_1 * overhead$ (7)

[0073] 对于等式 (7), 可以意识到更大的 ERC 片段数并不始终导致最佳的效率。相反, 对于小的文件, 应当使用小 ERC 片段数或甚至非 ERC。针对不同文件大小和 ERC 片段数计算了等式 (7) 中所需的总带宽, 并绘出了图 6 所示的曲线。在不同的 ERC 片段数之间的边界是适于特定 ERC 片段数的最优文件大小范围。例如, 图 6 的底部曲线示出了在其下应当使用非 ERC, 而在其上应当使用片段数 $k = 2$ 的 ERC 的文件大小边界。有趣的观察结果是文件大小边界相对地对对等体可用性并不敏感, 这极大简化了最优 ERC 片段参数的选择。一般而言, 对于比约 10KB 小的文件, 不应使用 ERC。对于片段数 $k = 2、4、6、8、16、32、128$ 和 256 的 ERC, 最合适的文件大小范围分别约为 10-33KB、33-100KB、100-310KB、310-950KB、950KB-2.9MB、2.9MB-8.9MB、8.9-26MB、> 26MB。

[0074] 2.4 自适应 ERC 模式

[0075] 自适应编码存储系统和方法自适应地选择合适的 ERC 片段数以可靠地在 P2P 网络中高效地存储内容。通过使用图 6 中建立的文件边界曲线, 该系统的一个实施例对于不同的文件大小选择使用非 ERC, 以及片段数 $k = 2、4、6、8、16、32、64、128、256$ 的 ERC。图 7 和图 8 示出了自适应 ERC 方法与固定参数的 ERC 的比较, 以及网络带宽使用中的区别, 其中对等体可靠性分别为 50% 和 99%。相比于使用 $k = 1$ (非 ERC)、8、32 和 256 的固定的 ERC 片段数, 对于 50% 的对等体可靠性, 自适应 ERC 方法可以将效率平均提高 61%、26%、25% 和

50%，而对于 99% 的对等体可靠性，平均提高 50%、18%、29% 和 57%。效率的提高是显著的。

[0076] 最一般而言，图 9 示出了自适应编码存储过程的一个实施例。如过程动作 902 所示，自适应编码存储系统计算对于不同片段数的最优文件大小边界。输入将要编码的给定文件大小的文件（过程动作 904）。如过程动作 906 所示，做出关于输入文件大小是否对应于非擦除编码 ($k = 1$) 的检查。如果输入文件大小不与 ERC 相对应，则不使用 ERC 来编码该文件（过程动作 908）。如果该文件大小对应于 ERC 文件大小范围，则使用 ERC 编码和对应于输入文件的文件大小的片段数来编码该文件，该片段数是对于该文件大小的最优片段数（过程动作 910）。

[0077] 此处所描述的自适应 ERC 过程的主要应用是 P2P 备份和还原。对等体可将文件备份到网络中的其他对等体并然后在文件丢失的情况下（例如，文件在计算机崩溃中丢失）通过网络中的对等体检索文件来还原这些文件。一般而言，图 10 示出了显示在 P2P 网络中如何利用自适应编码存储技术的示例性操作流程。应当注意，图 10 中由断开的线或虚线所表示的任何框和框之间的互连表示此处所描述的自适应编码存储系统的替换实施例，并且以下描述的这些替换实施例中的任一个或全部可以结合贯穿本文所描述的其它替换实施例来使用。

[0078] 具体地，如图 10 所示，在数据传输操作之前，诸如当期望将数据备份到网络中的对等体时，服务器 200 或对等体 220 编码 1000 将要传输到其他对等体以便存储的数据。自适应编码存储系统能够用多种传统编解码器中的任一个来运作，比如，MPEG 1/2/4、WMA、WMV 等等。此外，在编码过程 1000 期间，服务器 200 或对等体 220 还生成数据头，以及包含数据结构的同伴文件 (companion file)。

[0079] 如上所述，在一个实施例中，一旦编码 1000 了数据，该编码的数据分组就被拆分 1005 成多个固定大小的数据单元。此外，如同编码的数据一样，数据头和数据结构也被拆分 1005 成多个与用来拆分编码的数据分组相同的固定大小的数据单元。将该信息拆分 1005 成固定长度的数据单元允许对等体在数据传输操作之前预分配存储器块，由此避免在数据传输过程期间的计算上昂贵的存储器分配运算。此外，对较小的数据单元的使用允许储存数据的客户机或对等体对由每个对等体消耗的确切的带宽数量进行更精细的控制以在数据传输操作期间满足客户机数据单元请求。

[0080] 除了将编码的数据、数据头、以及数据结构拆分 1005 成更小的数据单元之外，如果利用擦除弹性编码，则还使用一额外的编码层以便在其中服务对等体固有地不可靠的典型 P2P 环境中提供增加的冗余。具体地，如上所述，在一个实施例中，如果确定擦除弹性编码适于数据文件，则数据单元进一步被分成多个数据块并且使用擦除弹性编码过程 1010 来编码文件。

[0081] 对这一编码 1010 的使用保证一个或多个对等体具有重构特定数据单元所必需的数据块，同时简化了客户机标识哪些对等体包含必要数据的需求。此外，在一个实施例中，每个服务对等体 220 所使用的擦除弹性编码密钥由服务器 200 自动地分配给每个对等体。然而，在另一个实施例中，每个服务对等体 220 简单地随机选择擦除弹性编码。这些密钥然后在每个对等体 220 最初由客户机 210 联系时由该客户机来检索。

[0082] 一旦最初已编码 1000 了数据文件，拆分 1005 成数据单元，并且可能进一步进行了

擦除编码 1010, 则所得数据单元或数据块然后就被分发 1015 到各对等体 220。该分发 1015 在将编码的数据块或分组简单地全部或部分提供给其中该数据块或分组被高速缓存或存储以备将来当希望检索该数据的客户机调用时进行数据传输的多个对等体的意义上可以是蓄意的。

[0083] 一旦数据已被分发 1015 到服务对等体 220, 客户机 210 然后在其希望从存储中检索该数据的情况下已准备好开始对那些对等体的数据请求。此外, 如上所述, 服务器 200 也可出于将数据传输到客户机 210 的目的充当对等体 220。

[0084] 此时, 客户机 210 通过首先检索可用的服务对等体 220 的列表来开始数据传输会话。该列表是直接从服务器 200、从对等体 220 中的一个、或通过使用用于标识潜在服务对等体的传统的分布式散列表 (DHT) 方法来检索的。一旦客户机 1010 已检索了对等体列表, 该客户机然后就连接到每个服务对等体 220 并且从每个对等体检索 1025 可用文件的列表。一旦客户机 210 已检索了每个对等体 220 的可用文件列表, 该客户机然后通过经由客户机和一个或多个对等体之间的网络连接向这些对等体请求对应于该信息的数据单元来检索 1035 将从这一个或多个对等体传输的数据的数据头和数据结构。

[0085] 数据头一般包含描述数据的全局信息, 例如, 数据中的通道数量、每个通道的属性和特性 (音频采样速率, 视频分辨率 / 帧速率)、所使用的编解码器、媒体的作者 / 版权所有等等。结果, 在数据传输会话开始时对数据头的检索允许客户机 220 设置或初始化 1040 必要的工具以在数据传输会话期间接收到随后接收到的分组之前解码 1065 这些分组。

[0086] 此外, 在检索 1035 了特定数据的数据结构之后, 客户机分析该数据结构并计算在数据传输过程期间将被请求的已传输数据的数据单元的数据单元 ID 1045。客户机 210 然后向对等体 220 中的一个或多个挨个请求那些数据单元 1050。

[0087] 最后, 一旦已经根据客户机 210 的请求 1050 检索了构成特定数据分组的所有数据单元, 就将那些数据分组重新装配 1055 成原始数据分组。然后解码 1060 重新装配的数据分组并且可在客户机 210 上还原 1065。

[0088] 3. 0P2P 存储 : 政策和设计策略

[0089] 除了基于将在 P2P 网络中存储的文件大小来调整 ERC 片段数之外, 还可提高效率。此处所描述的自适应编码存储系统的各实施例被设计成通过利用如下所述的某些策略来提高存储效率。可结合自适应编码存储系统或在任何 P2P 网络中利用这些策略。

[0090] 3. 1P2P 存储成本

[0091] 在本节中, 将在 P2P 网络中存储文件与直接在“六个 9”可靠服务器中存储文件相比较。可观察到, P2P 解决方案减少了服务器带宽和成本, 但要求对等体花费更多的带宽以将文件分发到 P2P 存储中。对网络带宽的总体使用在 P2P 解决方案中增加了。客户机上传带宽的增加可被认为是 P2P 存储系统的成本。对于不同对等体可靠性和文件大小的成本以表格形式示于表 1 中。

[0092] 表 1P2P 中增加的带宽使用的成本。

[0093]

	文件大小				
可靠性	10KB	100KB	1MB	10MB	100MB
10%	332.9	79.1	29.5	16.5	12.5
50%	51.0	12.11	4.34	2.23	1.56
99%	9.4	1.87	0.65	0.22	0.09

[0094] 可观察到,如果对等体可靠性很高且文件大小很大,则使用 P2P 存储的成本就很小。例如,将 100MB 的文件存储到具有 99%可靠性的对等体仅导致 9%的成本。然而,当对等体可靠性很低且文件大小很小时,该成本可能是相当大的。

[0095] 3. 2P2P 存储政策

[0096] 从表 1 中,可以导出以下使用 P2P 存储云的政策:

[0097] a) 应当使用不可靠的对等体来存储大文件,而使用可靠的对等体来存储小文件。如果将大文件分配给不可靠的对等体,而将小文件分配给可靠的对等体,则对于 P2P 系统的成本将更少。

[0098] b) 应当使用不可靠的对等体来存储静态文件,而使用可靠的对等体来存储动态文件。将那些不改变的文件称为静态的,而将那些一直改变的文件称为动态的。多个小静态文件可被捆绑成一个大静态文件并且存储在 P2P 存储云中。该同一策略对于动态文件并不有效,因为单个文件的改变需要更新整个组合的文件。

[0099] 该政策的推论是如果使用 P2P 网络来存储应用程序的状态、对等体状态信息等等,则应当将信息转向网络中最可靠的对等体。如果限制包含应用程序状态的文件只被放置在高可靠性对等体中(本质上,高可靠性对等体将形成构成扩展的 P2P 网络的核心子网),则可以极大地降低此重复率以及更新状态文件的成本,并且提高了效率。

[0100] c) 不可靠的对等体应被允许分发较少的文件,而可靠的对等体应被允许分发较多的文件。

[0101] d) 较小的文件应被分配较高的分发成本,而较大的文件应被分配较低的分发成本。

[0102] 政策 c) 和 d) 是对于 P2P 备份和检索应用的,其中对等体可以将内容分发到 P2P 存储云中,并且为其他对等体存储内容。平衡的 P2P 存储网络应当令每个对等体平衡其贡献和收益。在先前的工作中已经指出,带宽是 P2P 存储应用中主要资源。令对等体的贡献是其为其他对等体接收并存储的编码片段的数量。令对等体的收益是其分发到 P2P 云中的内容的数量。考虑到低可靠性导致更多的冗余数据存储,应当惩罚不可靠的对等体以使其将被允许分发较少的文件,并且奖励可靠的对等体以使其将被允许分发较多的文件。这一政策在 P2P 经济中具有积极的收益,因为它鼓励用户保持 P2P 应用程序在线,因此提高了 P2P 网络的总体可靠性并降低了所需的重复率。

[0103] 还可通过将高分发成本分配给小文件,从而要求对等体成比例地贡献更多,来惩罚小文件的分发;以及通过将低分发成本分配给大文件,从而令对等体成比例地贡献更少,来奖励大文件的分发。作为推论,P2P 备份应用程序应被设计为最小化备份频率。可以考虑

将多个改变捆绑成一大文件,并且仅一次,例如每天午夜将该文件更新到 P2P 存储云中,而不是在其改变之后立即更新该文件。

[0104] 图 11 示出了围绕上述政策设计的自适应编码存储系统和方法的一个实施例。如过程动作 1102 所示,确定在分布式或 P2P 网络中的每个对等体的可靠性。输入将要分发或存储的文件(过程动作 1104)。评估该文件的大小(过程动作 1106),并且基于等式(7)中的期望存储带宽将分发成本分配给该文件(过程动作 1108)。如果该文件是大文件,则可分配到较高的分发成本。如果该文件是小文件,则可将较低的分发成本分配给该文件。基于文件的大小,自适应编码存储系统将选择具有适当可靠性的对等体来存储该文件(过程动作 1110)。即,使用其可靠性低于给定阈值的对等体来存储和分发大文件,而使用其可靠性高于给定阈值的对等体来存储和分发小文件。

[0105] 图 12 示出了围绕上述政策设计的自适应编码存储系统和方法的另一个实施例。如过程动作 1202 所示,确定在分布式或 P2P 网络中的每个对等体的可靠性。输入将要分发或存储的文件(过程动作 1204)。将该文件与先前存储的同一文件相比较以确定该文件是静态的还是动态的(过程动作 1206)。在第一次存入该文件时,假设该文件是动态的。如果观察到对该文件的频繁改变,则该文件仍被指定为动态的。如果观察到该文件长时间没有改变,则该文件被指定为静态的。将动态文件存储在高可靠性对等体中(过程动作 1210)。(因此,首先,文件首先将被存储在服务器或高可靠性对等体中)。一旦观察到文件没有改变,并且变为静态的,则这些静态文件将被重新分发,并存储在低可靠性对等体中。

[0106] 应当注意,可以单独或组合地适用图 11 和 12 所示的各实施例以提高分布式或对等网络的总体效率和可靠性。

[0107] 3.3 具有服务器组件支持的 P2P 存储。

[0108] 如果使用服务器组件作为对 P2P 网络的补充,则可对于大的和静态文件使用 P2P 存储,而对于小的、动态文件使用服务器。因为消耗大多数服务器资源的是大文件,所以 P2P 存储是对服务器的很好的补充。

[0109] 如图 13 所示,自适应编码存储系统和过程的一个实施例利用了具有服务器支持的 P2P 备份。如图 13 所示,将网络中的动态文件备份到服务器(过程动作 1302)。客户机和/或服务器然后可以自动地检测那些不再改变并正转变为静态文件的动态文件(过程动作 1304、1306)。如过程动作 1308 所示,这些检测到的静态文件然后可一起被捆绑成一大文件,并用 ERC 分发到 P2P 存储云中(过程动作 1310)。这有效地增加了存储在 P2P 云中的文件的大小。结合片段数的 ERC,这可以提高效率。

[0110] 可以单独或与图 11 和 12 所示的各实施例组合地使用图 13 所示的实施例以提高分布式或对等网络的总体效率和可靠性。还应注意,该实施例可在使用或不使用擦除弹性编码的两种情况下使用。

[0111] 应当注意,可以按所需的任何组合使用上述替换实施例的任一个或全部以形成另外的混合实施例。尽管用对结构特征和/或方法动作专用的语言描述了本主题,但可以理解,所附权利要求书中定义的主题不必限于上述具体特征或动作。相反,上述具体特征和动作是作为实现权利要求的示例形式公开的。

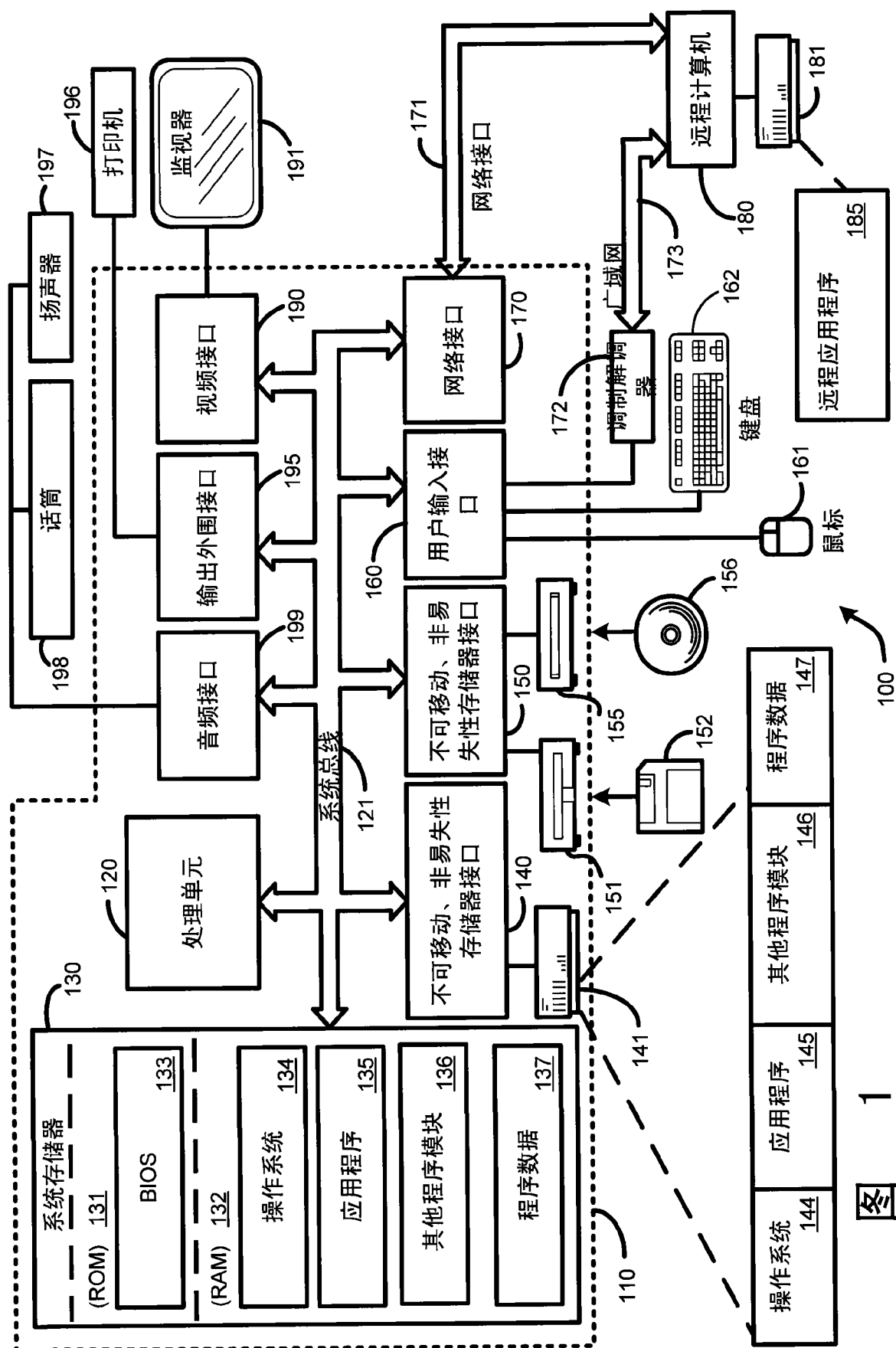


图 1

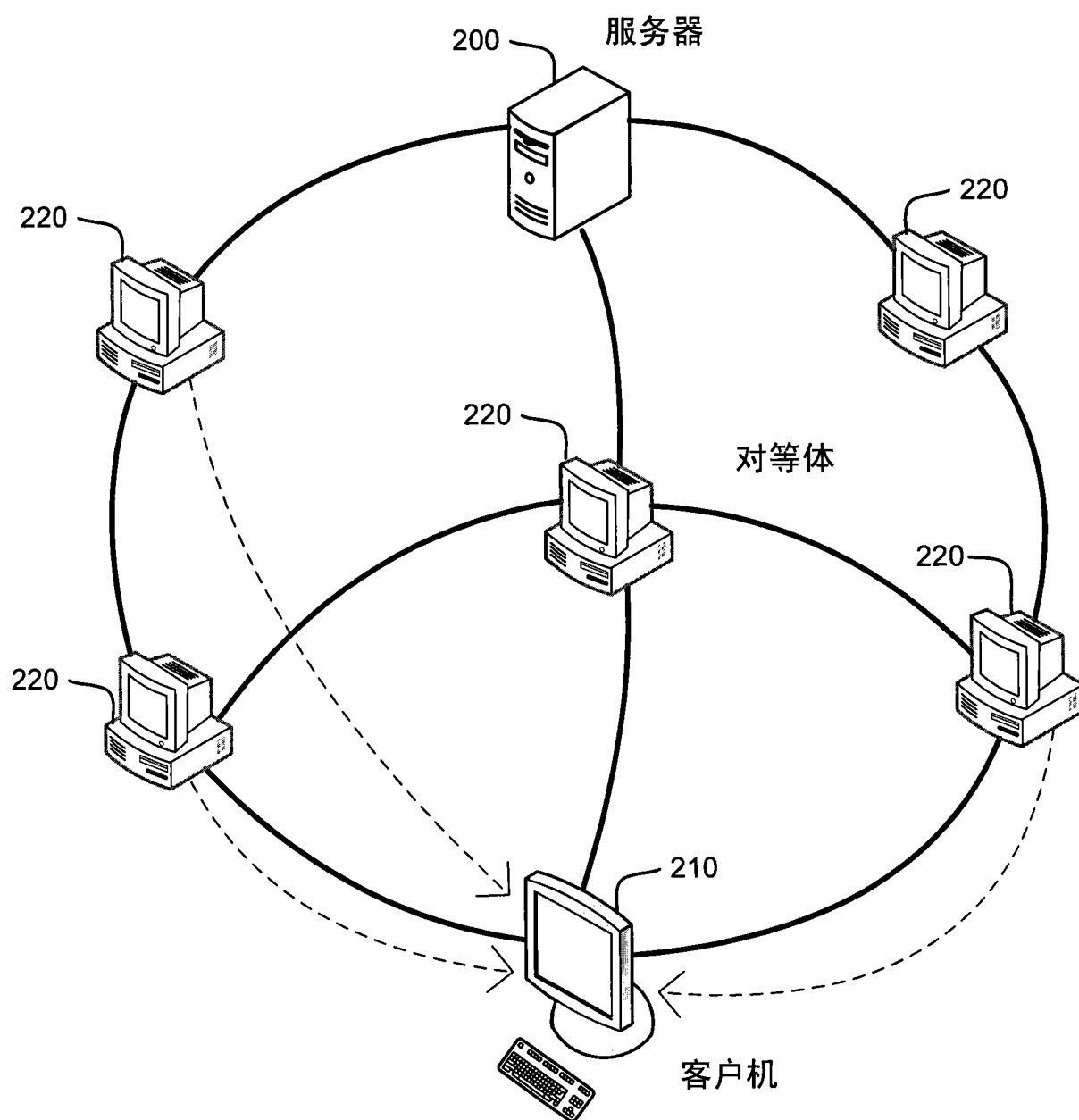


图 2

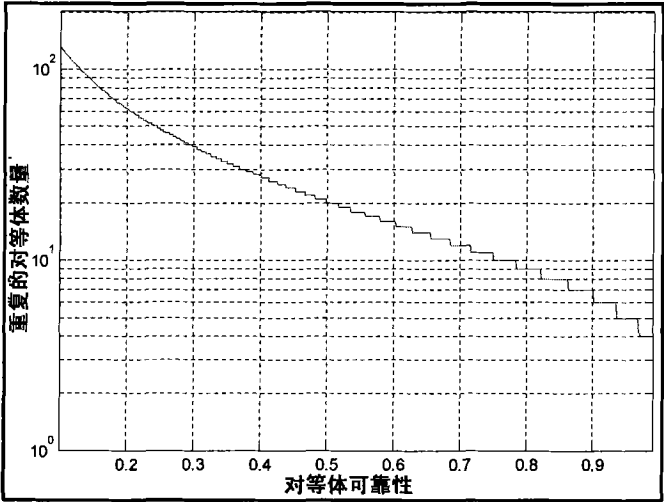


图 3

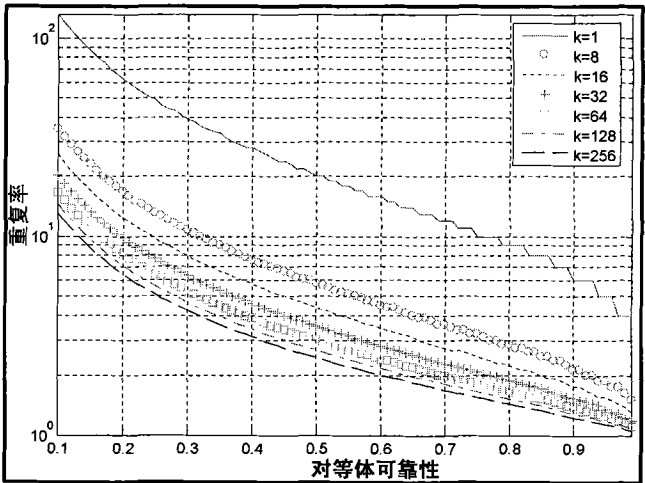


图 4

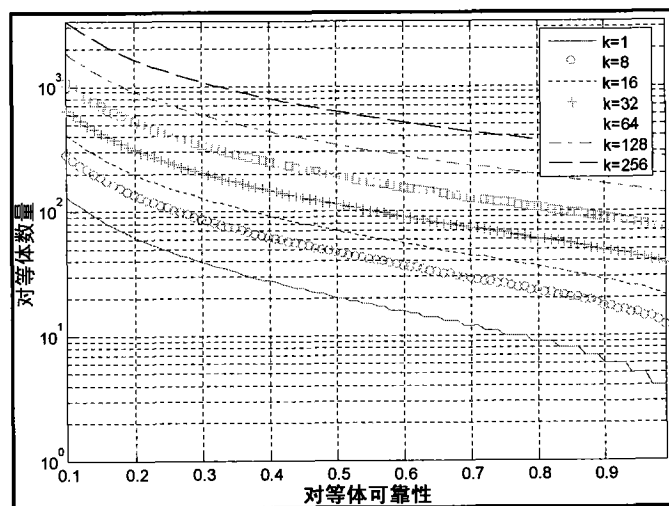


图 5

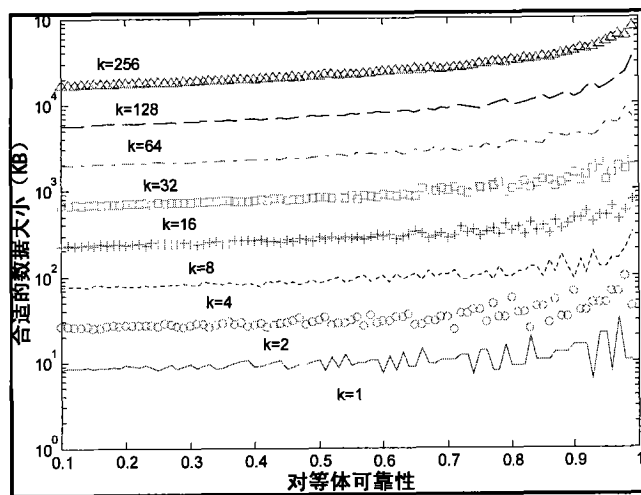


图 6

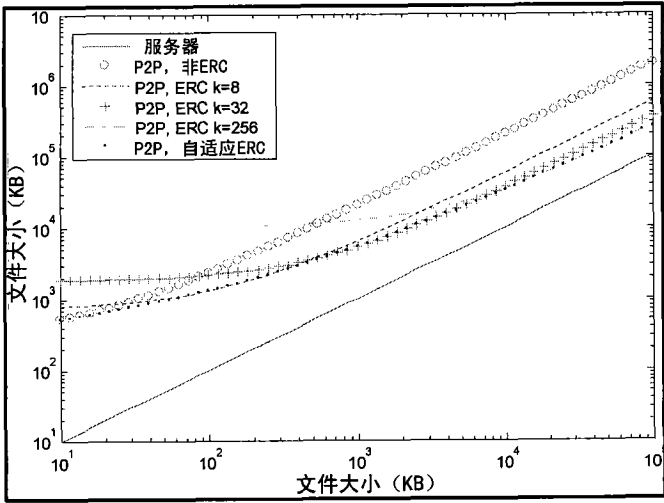


图 7

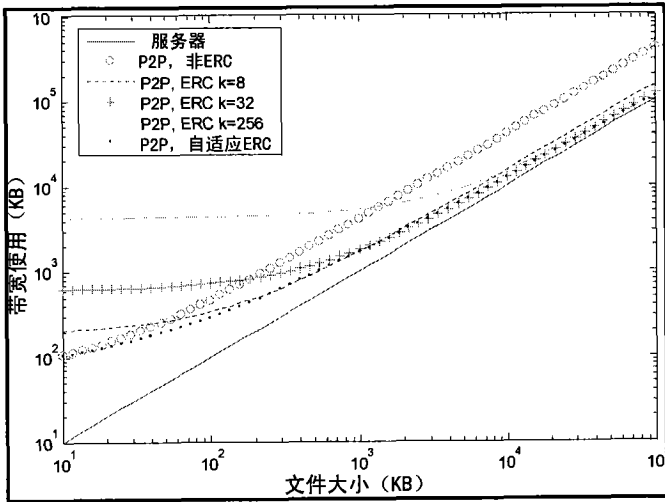


图 8

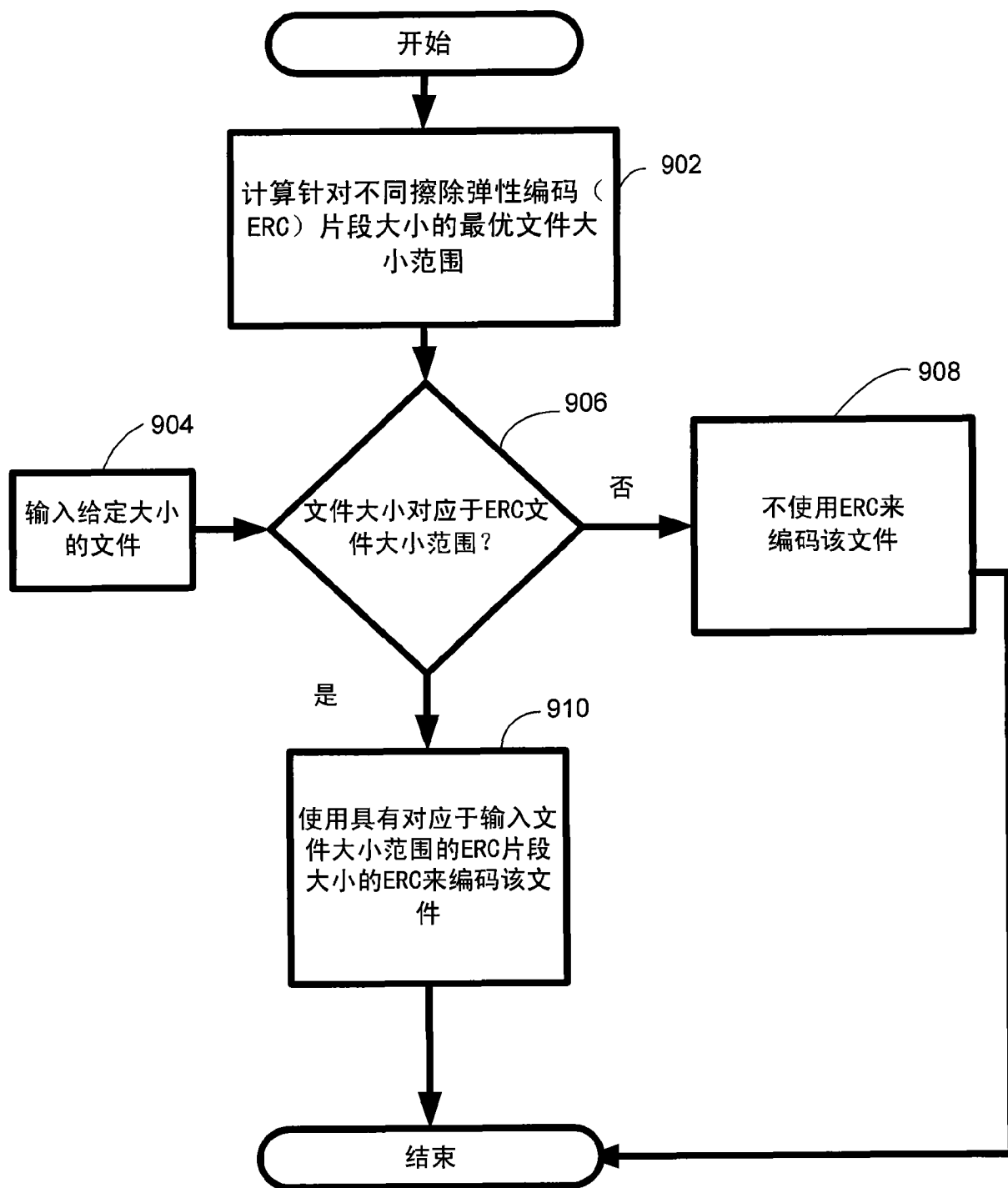
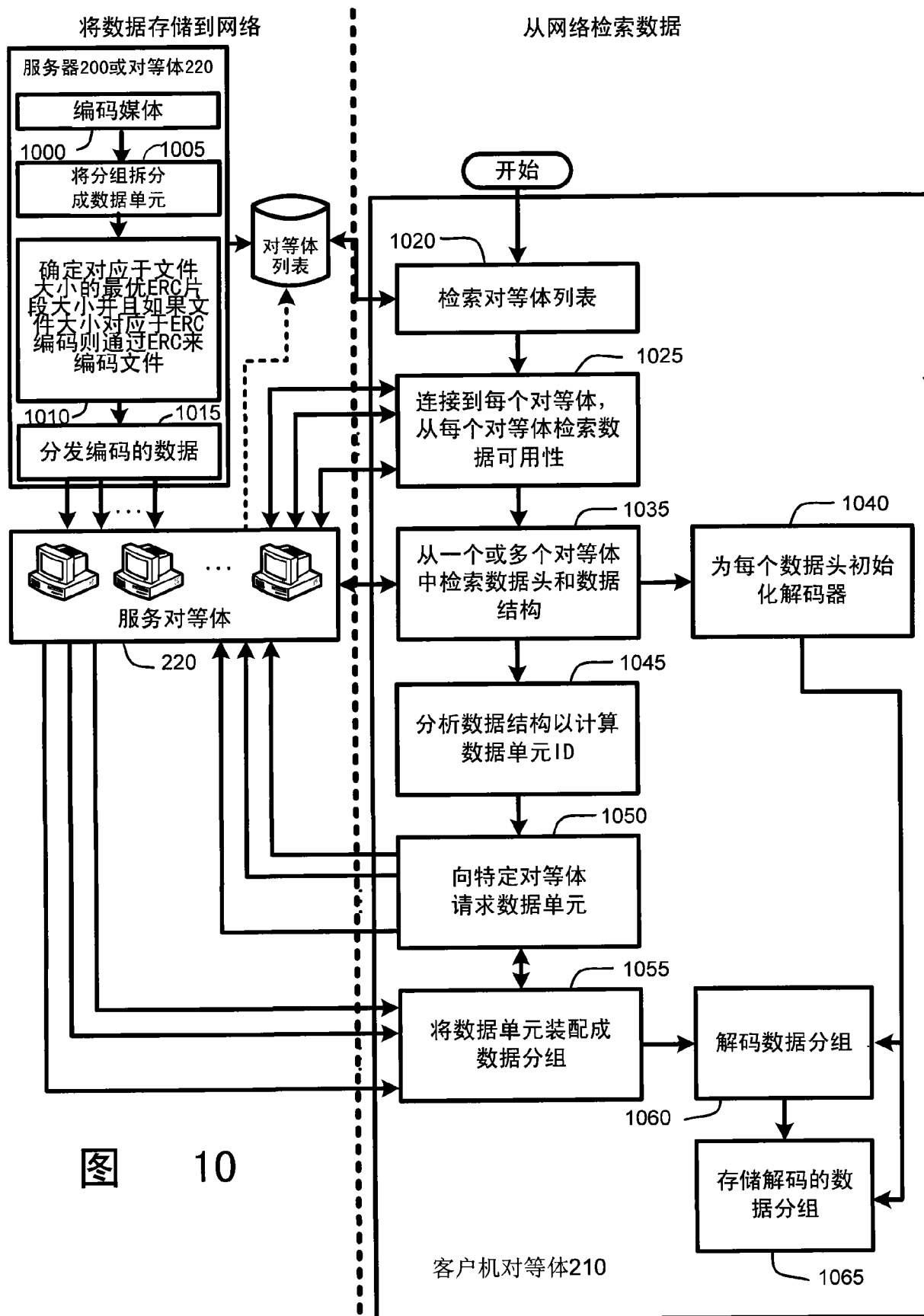


图 9



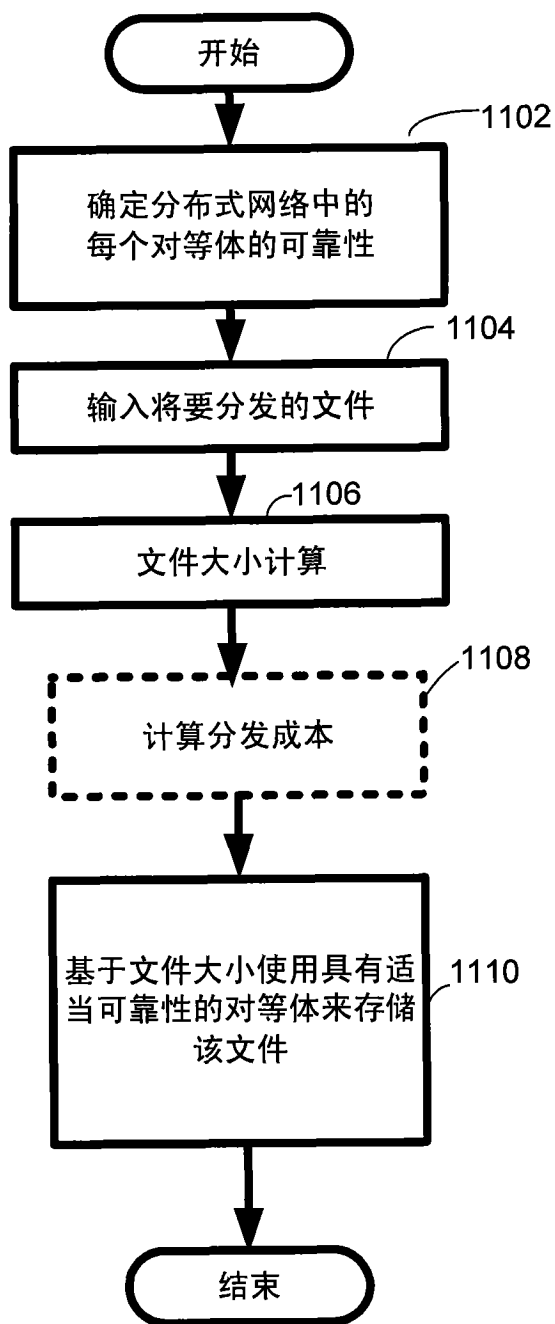


图 11

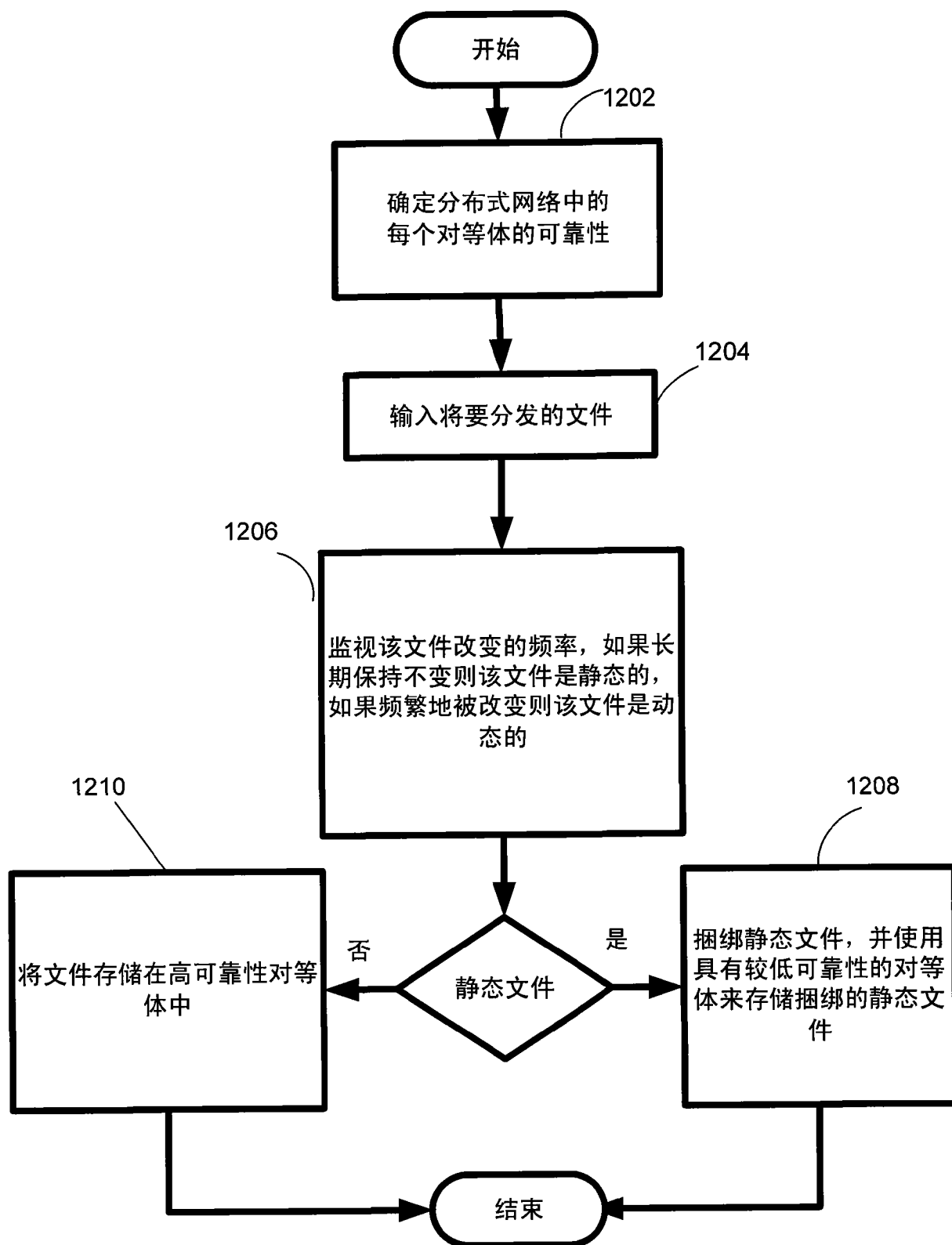


图 12

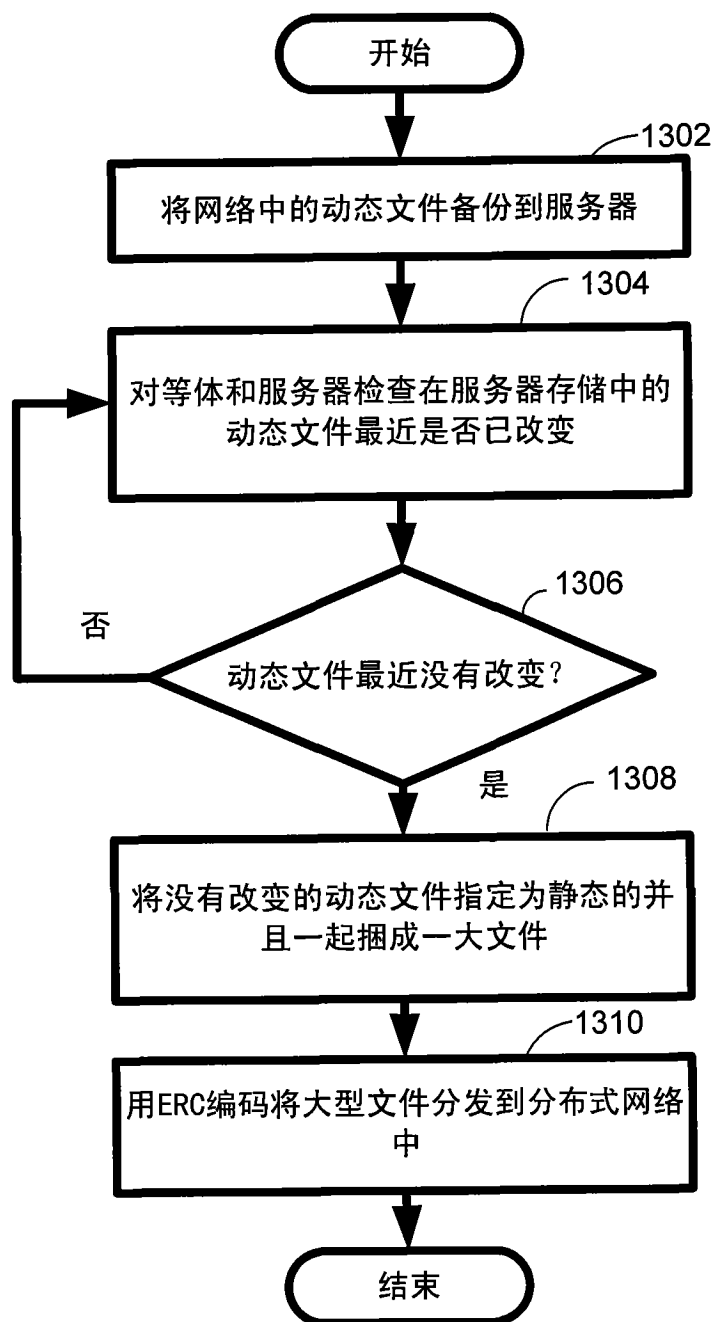


图 13