(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109804354 B
(45)授权公告日 2023.08.22

(21)申请号 201780062552.7
(22)申请日 2017.08.14
(65)同一申请的已公布的文献号
申请公布号 CN 109804354 A
(43)申请公布日 2019.05.24
(30)优先权数据
15/254,278 2016.09.01 US
(85)PCT国际申请进入国家阶段日
2019.04.10
(86)PCT国际申请的申请数据
PCT/US2017/046757 2017.08.14
(87)PCT国际申请的公布数据
WO2018/044538 EN 2018.03.08
(73)专利权人 甲骨文国际公司
地址 美国加利福尼亚
(72)发明人 M·杰斯沃 S·博塞 J·W·斯塔莫斯 A·R·道宁 D·辛格
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 周衡威
(51)Int.Cl.
G06F 9/54(2006.01)
(56)对比文件
CN 105393251 A,2016.03.09
CN 102377682 A,2012.03.14
CN 102035751 A,2011.04.27
CN 101311894 A,2008.11.26
CN 103179050 A,2013.06.26
US 2003110232 A1,2003.06.12
US 2006218560 A1,2006.09.28
审查员 李爽

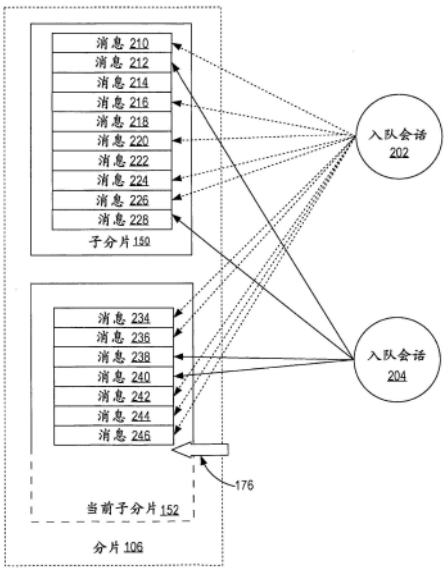
权利要求书3页 说明书18页 附图10页

(54)发明名称

用于消息队列的消息高速缓存管理

(57)摘要

提供了用于消息队列的消息高速缓存管理的方法和装置。将来自多个入队器的多个消息入队在包括一个或多个分片的队列中,每个分片包括一个或多个子分片。在存储器中维护消息高速缓存。将消息入队包括将消息入队在特定分片的当前子分片中,这包括将消息存储在与该特定分片的当前子分片对应的经高速缓存的子分片中。对于每个出队器-分片对,确定出队速率。生成估计访问时间数据,该估计访问时间数据包括基于出队器-分片对出队速率的、多个子分片中的每个子分片的最早估计访问时间。基于针对多个子分片的最早估计访问时间,确定一组子分片以存储为消息高速缓存中的经高速缓存的子分片。



CN 109804354 B

1.一种计算机实现的方法,用于消息队列的消息高速缓存管理,所述方法包括:

将来自多个入队器的多个消息入队在包括一个或多个分片的队列中,所述一个或多个分片中的每个分片包括一个或多个子分片;

在存储器中维护消息高速缓存,所述消息高速缓存被配置为存储多个经高速缓存的子分片;

其中,所述多个消息包括特定消息;

其中,将所述多个消息入队包括将所述特定消息入队在所述一个或多个分片中的特定分片的当前子分片中,其中,将所述特定消息入队在所述当前子分片中包括将所述特定消息存储在与所述特定分片的所述当前子分片对应的经高速缓存的子分片中;

其中,多个出队器-分片对中的每个出队器-分片对包括:(a)多个出队器中的从所述一个或多个分片中的相应分片进行出队的相应出队器;以及(b)所述相应分片;

至少通过针对所述多个出队器-分片对中的每个出队器-分片对确定由所述每个出队器-分片对的相应出队器对所述每个出队器-分片对的相应分片的出队速率,确定所述多个出队器-分片对的出队速率;

生成包括最早估计访问时间的估计访问时间数据,所述最早估计访问时间包括基于所述多个出队器-分片对的出队速率的、针对所述队列的多个子分片中的每个子分片的最早估计访问时间;

基于针对所述多个子分片的最早估计访问时间,确定所述队列的要存储为所述消息高速缓存中的经高速缓存的子分片的一组子分片;

将所述一组子分片维护为所述消息高速缓存中的经高速缓存的子分片;

当所述多个子分片中的第一子分片正被维护为所述消息高速缓存中的经高速缓存的子分片时,基于针对所述多个子分片的估计访问时间数据,确定所述第一子分片在特定时间帧内可能不会被访问;以及

响应于确定所述第一子分片在所述特定时间帧内可能不会被访问,从所述消息高速缓存中逐出所述第一子分片;

其中所述方法由一个或多个计算设备执行。

2.如权利要求1所述的计算机实现的方法,还包括:

在辅助存储装置中维护交换表;

当从所述消息高速缓存中逐出特定子分片时,将所述特定子分片的表示存储在所述交换表的行中;

当所述特定子分片被恢复到所述消息高速缓存时,至少部分地通过将所述特定子分片的所述表示从所述交换表读到所述消息高速缓存中来恢复所述特定子分片。

3.如权利要求1所述的计算机实现的方法,其中,每个子分片的最早估计访问时间是所述多个入队器中的任何入队器以及所述多个出队器中的任何出队器对于该子分片的最早估计访问时间。

4.如权利要求1所述的计算机实现的方法,还包括:

基于针对所述多个子分片的估计访问时间数据,生成子分片的有序访问列表;

其中,所述有序访问列表包括来自多个分片的子分片,所述子分片按估计访问时间排序而不考虑每个子分片所属的分片。

5.如权利要求4所述的计算机实现的方法,还包括:

利用特定出队器-分片对的更新后的出队速率,更新所存储的统计信息,所述统计信息指示所述一个或多个分片的出队器-分片对的出队速率;

在更新所存储的统计信息之后,至少部分地基于更新后的出队速率来重新计算所述有序访问列表。

6.如权利要求5所述的计算机实现的方法,其中:

所述特定出队器-分片对包括第一出队器和第一分片,所述第一分片包括第一子分片;并且

当所述第一出队器完成将所述第一子分片中的消息出队时或者当所述第一出队器开始处理所述第一分片的下一个子分片时,更新所述特定出队器-分片对的出队速率。

7.如权利要求5所述的计算机实现的方法,还包括:

由后台进程监视所述多个出队器的出队进度;

所述特定出队器-分片对包括第一出队器和第一分片,所述第一分片包括第一子分片;并且

其中,当所述后台进程确定所述第一出队器在处理所述第一子分片的时候已经拖延时,更新所述特定出队器-分片对的出队速率。

8.如权利要求1所述的计算机实现的方法,还包括:

为所述队列在辅助存储装置中维护队列表;

其中,将所述多个消息入队到所述当前子分片包括:通过将所述特定消息添加到所述队列表中的指派给所述当前子分片的分区内的所述队列表的行,来将所述特定消息存储在辅助存储装置中。

9.如权利要求1所述的计算机实现的方法,其中:

所述多个消息包括无序消息;

其中,生成包括最早估计访问时间的估计访问时间数据还基于所述无序消息的递送时间,所述最早估计访问时间包括针对所述队列的所述多个子分片中的每个子分片的最早估计访问时间。

10.如权利要求1所述的计算机实现的方法,在从所述消息高速缓存中逐出所述多个子分片中的第二子分片之后:

基于针对所述多个子分片的估计访问时间数据,确定所述第二子分片在第二时间帧内可能会被访问;

响应于确定所述第二子分片在所述第二时间帧内可能会被访问,预获取所述第二子分片以在所述消息高速缓存中恢复所述第二子分片。

11.一个或多个存储用于消息队列的消息高速缓存管理的指令的非瞬态计算机可读介质,所述指令在由一个或多个处理器执行时使得执行如权利要求1-10中任一项所述的计算机实现的方法。

12.一种用于消息队列的消息高速缓存管理的装置,包括用于执行如权利要求1-10中任一项所述的计算机实现的方法的部件。

13.一种用于消息队列的消息高速缓存管理的计算设备,包括:

一个或多个处理器;以及

存储器，耦合到所述一个或多个处理器并且包括存储于所述存储器上的指令，所述指令在由所述一个或多个处理器执行时，使得执行如权利要求1-10中任一项所述的计算机实现的方法。

用于消息队列的消息高速缓存管理

技术领域

[0001] 本文描述的实施例一般而言涉及队列，更具体而言，涉及用于消息队列的消息高速缓存管理的技术。

背景技术

[0002] 在许多应用中，在计算机系统上执行的一个进程必需与在同一个或其它计算机系统上执行的一个或多个其它进程通信。用于实施这些通信的机制因系统而异。促进各种系统中的进程到进程通信的一种机制是消息队列。进程通过将消息在消息队列中入队来将信息发送到其它进程。接收进程通过将消息从消息队列中出队来获得该信息。通常，这些消息以先进先出的方式被读取。消息队列的实施方案在美国专利No.7,181,482、美国专利No.7,185,033、美国专利No.7,185,034、美国专利No.7,203,706、美国专利No.7,779,418、美国专利No.7,818,386、美国专利No.7,680,793、美国专利No.6,058,389和美国专利No.8,397,244中描述，这些专利的内容通过引用整体并入本文。

[0003] 消息队列可以在存储器中或在辅助存储装置(诸如磁盘、光盘或固态驱动器，或任何其它持久的辅助存储装置)上实现。存储器中(in-memory)的消息队列允许队列操作在存储器中发生，从而减少I/O延迟。但是，存储器一般是更有限的资源。因此，不是总能假设消息队列可以完全在存储器中实现。

[0004] 由辅助存储装置支持的存储器中消息高速缓存可以被用于存储存储器中的消息队列中的至少一部分消息。例如，数据库支持的队列可以被架构为处理极大的队列，即使当聚合队列尺寸比可用存储器的尺寸大许多倍时。在数据库实现的消息队列中，入队进程使用与数据库的连接或入队会话来将消息入队，并且出队器使用出队会话来使消息出队。

[0005] 消息队列的常规实施方案不能很好地缩放。具体而言，随着出队会话的数量增加，对队列头部处的“热”消息的争用增加，从而降低了性能。此外，当入队会话和出队会话散布在多个系统上时，系统之间的网络和/或互连上的通信量会变得过多。

[0006] 分片(sharded)队列解决了这些问题中的一些问题。分片队列包括一个或多个分片。在每个分片内，消息基于入队时间被排序。但是，在分片之间不强制执行消息次序。通常，出队会话以先进先出次序使消息从每个分片出队。但是，在分片之间不强制执行出队次序。分片队列的实施方案在美国专利申请公开No.2014/0372486、美国专利申请公开No.2014/0372489和美国专利申请公开No.2014/0372702中描述，这些申请的内容通过引用整体并入本文。

[0007] 存在各种高速缓存算法用于选择要存储在存储器中的数据的子集。这些算法包括诸如先进先出(FIFO)和最近最少使用(LRU)之类的次优算法，以及诸如用于虚拟存储器交换的最优页面替换(OPT)之类的最优算法。但是，这些技术不能直接转移到其中在分片之间不强制执行入队次序或出队次序的分片队列，和/或不能直接转移到其中多个出队器可以潜在地从任何分片使消息出队的多出队器队列。例如，这种算法没有被设计来容纳多个入队器、多个出队器和/或多分片队列。此外，最优算法需要知道某些信息，而这些信息在具有

多个分片和多个出队器的队列中不容易确定。例如,OPT算法要求知道页面访问的顺序。

[0008] 因此,需要用于消息队列的最优消息高速缓存管理。

[0009] 本节中描述的方法是可以追求的方法,但不一定是先前已经构思或追求的方法。因此,除非另有说明,否则不应认为本节中描述的任何方法仅仅因为它们包含在本节中就有资格作为现有技术。

发明内容

[0010] 至少一个实施例涉及一种计算机实现的方法,包括:将来自多个入队器的多个消息入队在包括一个或多个分片的队列中,所述一个或多个分片中的每个分片包括一个或多个子分片;在存储器中维护消息高速缓存,所述消息高速缓存被配置为存储多个经高速缓存的子分片;其中,所述多个消息包括特定消息;其中,将所述多个消息入队包括将所述特定消息入队在所述一个或多个分片中的特定分片的当前子分片中,其中,将所述特定消息入队在所述当前子分片中包括将所述特定消息存储在与所述特定分片的所述当前子分片对应的经高速缓存的子分片中;其中,多个出队器-分片对中的每个出队器-分片对包括:(a)多个出队器中的从所述一个或多个分片中的相应分片进行出队的相应出队器;以及(b)所述相应分片;至少通过针对所述多个出队器-分片对中的每个出队器-分片对确定由所述每个出队器-分片对的相应出队器对所述每个出队器-分片对的相应分片的出队速率,确定所述多个出队器-分片对的出队速率;生成包括最早估计访问时间的估计访问时间数据,所述最早估计访问时间包括基于所述多个出队器-分片对的出队速率的、针对所述队列的多个子分片中的每个子分片的最早估计访问时间;基于针对所述多个子分片的最早估计访问时间,确定所述队列的要存储为所述消息高速缓存中的经高速缓存的子分片的一组子分片;将所述一组子分片维护为所述消息高速缓存中的经高速缓存的子分片;当所述多个子分片中的第一子分片正被维护为所述消息高速缓存中的经高速缓存的子分片时,基于针对所述多个子分片的估计访问时间数据,确定所述第一子分片在特定时间帧内可能不会被访问;以及响应于确定所述第一子分片在所述特定时间帧内可能不会被访问,从所述消息高速缓存中逐出所述第一子分片;其中所述方法由一个或多个计算设备执行。

附图说明

[0011] 在附图中:

[0012] 图1是描绘根据一个或多个实施例的示例分片队列的框图;

[0013] 图2是描绘根据一个或多个实施例的、在示例分片队列的特定分片中使消息入队的示例入队会话的框图;

[0014] 图3A是描绘根据一个或多个实施例的、使消息从示例分片队列出队的示例出队会话的示例的框图;

[0015] 图3B是描绘根据一个或多个实施例的、在时间帧内估计的出队进度的示例的框图;

[0016] 图4A是描绘根据一个或多个实施例的示例数据库系统的框图;

[0017] 图4B是描绘根据一个或多个实施例的示例估计访问时间数据的框图;

[0018] 图5是描绘根据一个或多个实施例的、用于示例分片队列的辅助存储装置中的示

例队列表的框图；

[0019] 图6是图示根据一个或多个实施例的、用于消息高速缓存管理的示例处理的流程图；以及

[0020] 图7是图示根据一个或多个实施例的、用于消息高速缓存管理的示例处理的流程图；以及

[0021] 图8图示了可以在其上实现一个或多个实施例的计算机系统。

具体实施方式

[0022] 在以下描述中，出于解释的目的，阐述了许多具体细节以便提供对实施例的透彻理解。但是，显而易见的是，可以在没有这些具体细节的情况下实践这些实施例。在其它实例中，以框图形式示出了众所周知的结构和设备，以避免不必要地模糊实施例。

[0023] 总体概述

[0024] 本文描述了用于消息队列的消息高速缓存管理的技术。分片队列包括一个或多个分片。在每个分片内，消息基于入队时间排序。但是，在分片之间不强制执行消息次序。分片队列中分片的数量可以在运行时期间改变。没有实现多个分片的队列是具有一个分片的队列。一般而言，每个出队器使来自分片队列的多个分片的消息出队，其中来自特定分片的所有消息按次序出队。在分片之间不强制执行出队次序。每个分片包括一个或多个有序的子分片。当出队器从特定分片中出队时，它处理这些有序的子分片以按照入队次序获得消息。

[0025] 收集并更新关于出队速率的统计信息，以确定用于多个子分片的估计访问时间数据。对于从分片队列中进行出队的每个出队器，将为每个出队器-分片对收集和更新统计信息。例如，对于特定的出队器，可以为分片队列的多个分片中的每个分片确定出队速率。

[0026] 通过使用用于出队器-分片对的出队速率，生成估计访问时间数据。估计访问时间数据包括关于对分片队列的未来访问的估计数据。例如，估计访问时间数据可以包括被调度为访问特定子分片的任何出队器对特定子分片的最早估计访问时间。

[0027] 用于多个子分片的估计访问时间数据被用于消息高速缓存管理。例如，估计访问时间数据可以被用于确定要在消息高速缓存中存储分片队列的哪些子分片。存储在消息高速缓存中的子分片在本文中称为“经高速缓存的子分片”。在一些实施例中，估计访问时间数据被用于确定应当从消息高速缓存中逐出经高速缓存的子分片和/或应当将被逐出的子分片存储在消息高速缓存中。

[0028] 这种方法通过消除对消息高速缓存中经高速缓存的子分片的一些不必要的逐出操作和/或从辅助存储装置的恢复操作来提高性能，诸如通过减少由辅助存储装置支持的队列的盘I/O。在关系数据库中，这种方法还通过减少用于检索存储在辅助存储装置中的数据的SQL执行开销来进一步提高性能。此外，这种方法更高效地使用消息高速缓存中的存储器。此外，这种方法减少了在出队器试图使消息从子分片出队时该子分片未存储在消息高速缓存中的情况的发生，从而避免了访问辅助存储装置中的消息的开销和延迟。在一些实施例中，这种方法基于每个分片上的每个出队器的未来出队速率和/或未来入队速率的预测而接近于用于虚拟存储器交换的最优页面替换(OPT)。

[0029] 队列和分片队列

[0030] 图1是描绘根据一个或多个实施例的示例分片队列的框图。分片队列100包括多个

分片102-106。虽然所示实施例示出了三个分片，但是用于实现分片队列的分片的实际数量可以根据实施方案变化。在一些实施例中，可以由管理员指定用于分片队列的分片的数量。可替代地和/或附加地，可以为分片队列100中的分片102-106的数量指定下限和/或上限。可替代地和/或附加地，实现分片队列100的系统可以确定分片的数量和/或动态地改变分片的数量。本文描述的实施例是关于具有一个或多个分片的分片队列来描述的，并且可以被推广到具有任意数量的分片的队列，包括具有一个分片的队列和/或没有被分片的队列。

[0031] 一个或多个入队器使消息入队在分片队列100中。如本文所使用的，术语“消息”是指要经由队列传送的任何数据。在一些实施例中，消息包括有效载荷和元数据的集合。分片102-106为同一个分片队列100存储不同的消息集合。在每个分片102-106内，基于入队时间对消息进行排序。但是，在分片102-106之间不强制执行消息次序。为了以入队次序存储项目，允许从存储的数据重建入队次序的任何存储手段都是足够的。

[0032] 在一些实施例中，来自特定入队器的所有消息在分片队列的特定分片中被入队。如本文所使用的，“入队亲和性”是指该特定分片与该特定入队器之间的关系。入队亲和性确保在没有失败的情况下满足会话排序要求，因为每个出队器都会看到每个入队器以正确的次序入队的消息。例如，当一个或多个入队器的第一集合将消息入队到分片队列100中时，消息总是被入队到分片102中；当一个或多个入队器的第二集合将消息入队到分片队列100中时，消息总是被入队到分片104中；当一个或多个入队器的第三集合将消息入队到分片队列100中时，消息总是被入队到分片106中。

[0033] 在一些实施例中，分片队列的分片102-106彼此分开维护。例如，分片102-106可以由多实例数据库的多个数据库服务器实例、其它服务器应用实例和/或计算单元维护。在一些实施例中，多实例数据库的每个数据库服务器实例维护分片队列中的单个分片102-106。可替代地和/或附加地，多实例数据库的一个或多个实例可以维护分片队列100中的多个分片102-106。可替代地和/或附加地，多个分片102-106可以由单个服务器、应用和/或计算单元维护。如本文所使用的，术语“服务器”是指集成软件部件和计算资源(诸如存储器、节点和节点上用于执行集成软件部件的进程)的分配的组合，其中软件和计算资源的组合专用于代表服务器的客户端提供特定类型的功能。

[0034] 子分片

[0035] 每个分片102-106可以被划分为一个或多个子分片110-170。如本文所使用的，术语“子分片”是指分片中的一个或多个相邻消息的序列，其中邻接性由消息入队时间确定。子分片包括在特定分片中入队的消息子集合。在子分片内，消息以入队次序存储。例如，可以基于消息被任何入队器入队到分片102中的入队时间来按次序检索子分片110内的所有消息。分片的子分片本身也按入队时间排序。例如，子分片110中的所有消息在子分片112中的消息之前被入队到分片102中，并且子分片112中的所有消息在子分片114中的消息之前被入队到分片102中，依此类推。在一些实施例中，本文描述的技术在没有被分片的队列上实现，该队列可以被视为具有单个分片的队列。在这种情况下，整个队列被视为包含多个子分片的单个分片。

[0036] 出队器可以基于子分片110-170的排序和每个子分片110-170内消息的排序，以入队时间次序访问每个分片中的消息。当特定的入队器仅将消息排入特定的分片时，出队器可以按次序将来消息从该特定的入队器出队，从而维持会话排序。

[0037] 当新消息被入队时，该新消息被添加到队列分片的分片尾部。在一些实施例中，分片尾部引用172-176识别相应分片102-106的队列分片尾部。例如，分片尾部引用可以是指向(例如，易失性存储器和/或辅助存储装置中的)当前子分片122、138和152的指针。如本文所使用的，术语“当前子分片”是指包括相应分片102-106的分片尾部的子分片。在当前子分片变满时，针对该分片的后续消息被入队到被分配给该分片的下一个子分片中。例如，当分片102的当前子分片122满时，子分片124变为当前子分片，并且针对分片102的后续消息被存储在子分片124中。在一些实施例中，当特定分片102与可以将消息入队到特定分片102中的一个或多个活动入队器相关联时，特定分片102的当前子分片122总是高速缓存在易失性存储器中，以便于入队操作和出队操作。

[0038] 随着附加消息在对应的分片102-106中入队，在稍后的时间点生成子分片124-128、140-148和154-170。在一些实施例中，可以在当前子分片变满并且需要下一个当前子分片时预分配和/或分配用于子分片的易失性存储器和/或辅助存储装置。子分片110-120、130-136和150用实线示出，以指示这些子分片已满。当前子分片122、138、152部分地用实线示出并且部分地用虚线示出，以指示这些子分片是部分满的。未来的子分片124-128、140-148和154-170用虚线示出，以指示这些子分片不包含消息。

[0039] 在逻辑级别，分片队列、队列中的每个分片以及队列的分片中的每个子分片各自与消息集合对应。当在计算机系统上实现时，生成对应的数据结构并将其存储在存储器和/或辅助存储装置中，如将在下文中更详细描述的那样。如本文所使用的，取决于术语在其中出现的上下文，术语“子分片”可以指消息的逻辑集合、存储在易失性存储器中的消息集合(例如“经高速缓存的子分片”)，和/或存储在辅助存储装置中的消息集合(例如，队列表中的行集合)。如本文所使用的，取决于术语在其中出现的上下文，术语“分片”可以指消息和/或子分片的逻辑集合、存储在易失性存储器中的消息集合(例如，一个或多个经高速缓存的子分片)和/或存储在辅助存储装置中的消息集合(例如，队列表中的行集合)。

[0040] 消息高速缓存

[0041] 分片队列100可以在易失性存储器和辅助存储装置两者中实现。如本文所使用的，术语“经高速缓存的子分片”是指存储在存储器中(诸如在消息高速缓存中)的子分片的表示。经高速缓存的子分片的表示也可以存储在辅助存储装置中，从而使得经高速缓存的子分片持久化。在一些情况下，诸如当分片队列100不完全适配到消息高速缓存中时，一个或多个子分片的表示可以仅存储在辅助存储装置中。当消息被入队到子分片中时，该消息被存储在易失性存储器中的子分片的表示和辅助存储装置中的子分片的表示当中的至少一者中。

[0042] 消息高速缓存提供到所有排队功能(包括入队操作、出队操作和通知操作)的快速存储器内访问路径。消息高速缓存可以存储来自一个或多个不同队列(包括一个或多个不同的分片队列)的消息。当对经高速缓存的子分片执行队列操作时，入队器和出队器不需要对辅助存储装置中的结构(诸如数据库表)进行扫描或排序。与I/O操作相比，存储器中的队列操作没有盘延迟。在一些实施例中，当消息高速缓存中有足够空间时，具有当前活动性的所有子分片存储在消息高速缓存中，以避免昂贵的SQL操作。例如，在一些实施例中，每个分片102-106的具有现有入队器的当前子分片122、138和152始终是经高速缓存的子分片。

[0043] 示例系统体系架构

[0044] 图4A是描绘根据一个或多个实施例的示例数据库系统的框图。数据库系统400包括数据库服务器412。数据库服务器管理并促进对一个或多个数据库的访问，从而处理客户端访问一个或多个数据库的请求。数据库服务器412在辅助存储装置402中管理与所存储的数据库对应的数据文件404。数据库服务器412还在辅助存储装置402中维护与分片队列100对应的持久数据(诸如队列表406)。数据库服务器412还在易失性存储器410中维护消息高速缓存414。在一些实施例中，消息高速缓存414被维护在数据库系统400的系统全局区域(SGA)中，该系统全局区域包括由数据库服务器412的所有进程共享的易失性存储器410。

[0045] 数据库服务器412在消息高速缓存414中维护并管理多个经高速缓存的子分片，以促进由入队会话416和出队会话418进行的对于分片队列100的存储器中队列操作。如本文所使用的，术语“会话”是指与数据库的连接，这可以包括数据库服务器412的一个或多个进程和/或一个或多个客户端进程。虽然关于入队会话和出队会话描述了一些实施例，但是本描述也适用于分片队列的入队器和出队器，无论入队器和/或出队器是否使用入队会话和/或出队会话来实施队列操作。

[0046] 入队器

[0047] 入队会话是允许入队器访问分片队列100的连接。例如，诸如进程之类的入队器可以经由入队会话将消息入队到分片队列100的特定分片中。如本文所使用的，术语“进程”是指在计算机中运行的程序指令集合的实例。进程可以具有虚拟地址空间、可执行代码、系统对象的打开句柄、安全上下文、唯一进程标识符、环境变量、优先级类、最小和最大工作集尺寸，和/或一个或多个执行线程。

[0048] 图2是描绘根据一个或多个实施例的、在示例分片队列的特定分片中使消息入队的示例入队会话的框图。入队会话202-204被指派给分片队列100的分片106(例如，与分片106具有入队亲和性)。即，通过入队会话202-204入队到分片队列100中的所有消息被入队到分片队列100的分片106中。

[0049] 入队会话202-204通过在分片106的当前子分片152的队列分片尾部176处添加消息210-246来使消息入队。分片106中的消息210-246以相对于指派给分片106的任何入队会话202-204的入队次序被存储。例如，由入队会话202入队的消息210在由入队会话204入队的消息212之前被入队。消息210、216、220、224、226、234、236、242-246由入队会话202中入队，消息212、228、238和240由入队会话204入队，并且消息214、218和222由另一个入队会话入队。

[0050] 出队器

[0051] 如本文所使用的，术语“出队器”是指消费来自分片队列的消息的任何实体。例如，出队器可以是使消息从分片队列出队的进程。为了消费消息，单个出队器可以使用任意数量的出队会话来消费来自单个队列的消息。出队会话是允许出队器访问分片队列的连接。在一些实施例中，当出队器具有多个出队会话时，多个出队会话必须协调与其它出队会话的消费，使得相同的消息不会被相同的出队器消费多于一次。

[0052] 一般而言，分片队列的任何分片102-106可以潜在地包含必须由分片队列100的任何出队器消费的消息。因此分片队列100的每个出队器一般处理分片队列的每个分片102-106，以使消息从分片队列100出队。在一些情况下，特定的出队会话可以使消息从分片队列的分片子集出队，这在美国专利申请公开No.2014/0372486、美国专利申请公开No.2014/

0372489和美国专利申请公开No.2014/0372702中更详细地描述。

[0053] 图3A是描绘根据一个或多个实施例的、使信息从示例分片队列出队的示例出队会话的框图。分片队列可以具有一个出队器或多个出队器。在所示实施例中，分片队列100具有两个出队器。出队会话380-382各自与分片队列100的出队器对应。每个出队会话380-382使消息从每个分片102-106出队。

[0054] 当分片队列具有多个出队器时，这些出队器可以彼此独立地起作用，并且可以在分片队列100中的每个分片102-106中的不同位置处进行出队。在一些实施例中，队列会话380-382使来自一个分片102的消息子集以相对于分片102的入队次序出队，然后切换到另一个分片104，以使来自该另一个分片104的消息以相对于该另一个分片104的入队次序出队。出队会话380-382访问分片队列100的分片102-106的次序可以基于各种因素确定。每个出队会话380-382所遵循的次序可以随时间相同和/或不同，并且可以与其它出队会话380-382相同和/或不同。此外，在单轮期间从特定分片102-108处理的消息和/或子分片的数量可以相同和/或不同，并且可以是自适应的。用于访问分片102-106的简单方案是轮询(round-robin)方案。

[0055] 在一些实施例中，每个出队会话380-382为每个分片102-106保持当前出队位置302-316以跟踪每个分片102-106中的出队进度。在所示实施例中表示的时间处，当前出队位置302指示出队会话380当前正在处理分片102的子分片114，而当前出队位置304指示出队会话382当前正在处理分片104的子分片138。在本文描述的一个或多个示例中，当前出队位置被示出为识别对应分片中的子分片的细节水平。在一些实施例中，用于出队器(或出队会话)的在分片上的当前出队位置包括子分片内的特定消息偏移或对子分片内的特定消息的另一种引用。

[0056] 当前出队位置310指示出队会话380将处理的分片106中的下一个子分片是子分片150。当前出队位置312指示出队会话380将处理的分片104中的下一个子分片是子分片132。当前出队位置314指示出队会话382将处理的分片106中的下一个子分片是子分片152。当前出队位置316指示出队会话382将处理的分片102中的下一个子分片是子分片118。

[0057] 在一些实施例中，每个出队会话380在继续到当前分片中的下一个子分片或者不同的分片中的下一个子分片之前对子分片整体进行完全地处理。例如，出队会话380将在继续到同一分片102的子分片116、分片104的子分片132或分片106的子分片150之前完全地完成处理子分片114。出队会话382将在继续到同一分片104的子分片140、分片106的子分片152或分片102的子分片118之前完全地完成处理子分片138。

[0058] 队列表

[0059] 图5是描绘根据一个或多个实施例的、用于示例分片队列的辅助存储装置中的示例队列表的框图。在分片队列100中入队的消息持久地存储在辅助存储装置402中的队列表406中。

[0060] 在一些实施例中，给定分片的每个子分片被指派一个或多个已指派给该给定分片的队列表分区510-552。例如，分片102被指派队列表406的一组分区510-512，并且其子分片102-122被指派到与分片102对应的队列表分区510-512；分片104被指派队列表406的一组分区530-532，并且其子分片132-138被指派到与分片104对应的队列表分区530-532；并且分片106被指派队列表406的一组分区550-552，并且其子分片150-152被指派到与分片106

对应的队列表分区550-552。

[0061] 在一些实施例中，可以将单个队列表分区指派给多个子分片。在替代实施例中，分片被划分为子分片而不考虑队列表300的分区。因此，子分片与队列表分区之间的关系可以是一对多、一对一、多对一、或者根本没有特别的关系。每个分片使用的分区的数量可以基于各种因素而变化，包括入队器将消息入队到每个分片的速率以及出队器从每个分片使消息出队的速率。因此，任何给定分片中的分区的数量可以随时间而变化，当入队器用完新消息的存储空间时会添加新分区，而当出队器完成分区中所有消息的出队时，该分区将被丢弃。

[0062] 在一些实施例中，对队列表406执行插入操作(诸如SQL INSERT)以将持久消息入队到队列表406中。在一些实施例中，对队列表406执行选择操作(诸如SQL SELECT)，以使消息从队列表406出队。可替代地，在一些实施例中，出队总是和/或主要从消息高速缓存414执行。

[0063] 消息高速缓存管理

[0064] 执行消息高速缓存管理，以确定要在消息高速缓存中维护的一组子分片。在一些实施例中，消息高速缓存管理至少部分地由一个或多个后台进程(诸如数据库服务器412的守护进程)执行。后面将更详细地描述示例后台进程。

[0065] 在理想情况下，所有出队器都跟上入队器，并且出队器需要处理的任何子分片将适配消息高速缓存。但是，当出队器跟不上入队器时，未处理的子分片的数量增加。出队会话418可能落后，使得它们必须从特定分片的除当前子分片之外的子分片进行出队，该当前子分片可能已经在消息高速缓存中以便于入队操作。

[0066] 例如，参考图3A，出队会话380在分片102、104和106中落后，并且出队会话382在分片102中落后。出队会话382在分片104中是当前的；如出队会话382的对于分片104的当前出队位置304所指示的那样，出队会话382当前在分片104的当前子分片138处从分片104进行出队。出队会话382在分片106中也是当前的；如出队会话382的对于分片106的当前出队位置314所指示的那样，出队会话382将在出队会话382访问分片106以使消息出队时，在分片106的当前子分片152处开始从分片106进行出队。

[0067] 为了解决这个问题，实现了在消息高速缓存中逐出和恢复子分片的机制，这将在下文中更详细地描述。如本文所使用的，术语“逐出”是指将经高速缓存的子分片从消息高速缓存移动到辅助存储装置。在子分片被逐出之后，该子分片不再是经高速缓存的子分片。如本文所使用的，术语“恢复”是指将子分片从辅助存储装置移到消息高速缓存。在恢复子分片之后，该子分片成为经高速缓存的子分片，并且可以在存储器中对经高速缓存的子分片执行队列操作。逐出操作和恢复操作允许在存储器中执行队列操作，而不会基于消息高速缓存的尺寸限制分片队列的尺寸。消息高速缓存管理涉及确定何时将具体的子分片从消息高速缓存414逐出到辅助存储装置402以及何时将具体的子分片从辅助存储装置402恢复到消息高速缓存414。

[0068] 出队速率

[0069] 在一些实施例中，基于一个或多个出队速率来执行消息高速缓存管理。图3B是描绘根据一个或多个实施例的、时间帧内的估计的出队进度的示例的框图。针对一个或多个出队器-分片对来确定出队速率。如本文所使用的，术语“出队器-分片对”是指代多个出队

器中的单个出队器和分片队列的单个分片。在出队器具有多于一个出队会话的情况下，出队器-分片对可以指代出队器的多个出队会话，或者出队器的特定出队会话。用于分片队列100的出队器-分片对是<380:102>、<380:104>、<380:106>、<382:102>、<382:104>和<382:106>。在图3B中，示出了出队器-分片对<380:102>和<382:102>。

[0070] 如本文所使用的，术语“出队速率”是指指示消息出队的速率的任何量化值。例如，出队速率可以是指示每单位时间内消息320-348的数量或者每个消息320-348的时间量的任何量化值。在一些实施例中，出队速率是指示每单位时间内子分片114-118的数量和/或每个子分片114-118的时间量的量化值。

[0071] 出队器-分片对的出队速率可以是指示消息通过特定出队器和/或出队会话从分片队列的特定分片出队的速率的任何量化值。每个出队器-分片对可以具有不同的出队速率。例如，出队会话380可以以第一出队速率从分片102进行出队，以第二出队速率从分片104进行出队，并且以第三出队速率从分片106进行出队。同样，出队会话382可以以第四出队速率从分片102进行出队，以第五出队速率从分片104进行出队，并且以第六出队速率从分片106进行出队。

[0072] 出队器-分片对的出队速率可以基于一段时间内的出队统计信息。例如，出队速率可以基于出队器从特定分片的历史出队速率。在一些实施例中，出队器-分片对的出队速率基于在一段时间内出队的消息的数量和/或处理的子分片的数量。出队器-分片对<380:102>的出队速率是$DR_{380,102}$，出队器-分片对<382:102>的出队速率是$DR_{382,102}$。

[0073] 在一些实施例中，可以对一个或多个出队统计信息进行加权，以确定出队器-分片对的出队速率。例如，为了计算出队速率，针对特定出队器-分片对收集的更近的出队统计信息可以比针对该特定出队器-分片对收集的不太近的出队统计信息更重地加权。例如，出队速率可以基于出队器-分片对的最近出队统计信息的加权和。可替代地和/或附加地，可以基于随时间的历史模式(例如，一天中的时间、日期、星期几、星期、月、年或任何其它历史模式)、系统的历史工作负载统计信息、一个或多个出队器的历史工作负载统计信息、管理员和/或用户输入、以及相对于出队消息的速率具有预测性的值的任何其它因素来确定出队速率。

[0074] 速率计算和重新计算

[0075] 在操作中，入队会话和出队会话的性能可以随时间而变化。例如，网络延迟、资源不可用性、工作负载改变以及其它因素可能会造成特定出队器-分片对的出队速率改变。因此，在操作期间持续地维护和更新统计信息。

[0076] 在一些实施例中，可以(诸如基于一段时间，或者基于处理的消息和/或子分片的数量)周期性地更新出队器-分片对的出队速率。在一些实施例中，每个出队会话被配置为在对子分片整体进行完全地处理之后继续到另一个子分片。

[0077] 在一些实施例中，当出队会话开始对子分片进行出队、完成子分片出队和/或以其它方式从子分片过渡到另一个子分片时(包括当出队会话从一个分片中的子分片过渡到不同分片中的子分片时)，进行对出队器-分片对出队速率的更新。在一些实施例中，当出队会话380-382开始或完成处理指定分片中的子分片(或设定数量的子分片)时，出队会话380-382或数据库服务器412的另一个元件更新出队器-分片对的出队速率和/或触发对出队器-分片对的出队速率的更新。

[0078] 估计访问时间数据

[0079] 给定每个分片中的每个出队器的当前出队位置302-316以及每个出队器-分片对的对应出队速率，可以生成估计访问时间数据。如本文所使用的，术语“估计访问时间数据”是指与将要访问一个或多个子分片的时间(诸如将对一个或多个子分片执行入队和/或出队操作的时间)的预测相关的任何数据。在一些实施例中，除了测量和预测每个出队器-分片对的出队速率之外，系统还测量并预测每个分片的入队速率。估计的入队速率使系统确保在不久的将来，对于入队器而言，在消息高速缓存中有足够的可用空间。

[0080] 参考图3B，基于出队器-分片出队速率$DR_{380,102}$和出队会话380在分片102中的当前出队位置302，估计出队会话380在未来时间T_f的未来出队位置350将在子分片114中。基于出队器-分片出队速率$DR_{382,102}$和出队会话382在分片102中的当前出队位置316，估计出队会话382在T_f处的未来出队位置352将在子分片122中。未来出队位置350-352是估计访问时间数据的示例。

[0081] 图4B是描绘根据一个或多个实施例的示例估计访问时间数据的框图。估计访问时间数据450包括多个子分片的多个最早估计访问时间。最早估计访问时间是任何出队器和/或出队会话380-382被估计为访问特定子分片的最早时间。例如，如图3B中所示，两个出队会话380-382都将在未来时间处理子分片120。给定出队器-分片对<380:102>的出队速率$DR_{380,102}$和出队会话380在分片102中的当前出队位置302，可以估计出队会话382将在第一时间(例如，T_1)处访问子分片120。

[0082] 给定出队器-分片对<382:102>的出队速率$DR_{382,102}$和出队会话382在分片102中的当前出队位置316，可以估计出队会话382将在第二时间(例如，T_2)处访问子分片120。因此，如果未来没有其它出队会话将访问子分片120，那么子分片120的最早估计访问时间Tc+d是T_1和T_2当中的最早时间。

[0083] 在一些实施例中，当一个或多个入队会话202-204正将消息入队到分片106的当前子分片152中时，当前子分片152被认为是当前被访问的。当前被访问的子分片可以被指派当前时间T_c作为最早估计访问时间。

[0084] 有序访问列表

[0085] 在一些实施例中，估计访问时间数据包括子分片的有序访问列表。有序访问列表包括指示所估计的未来子分片访问的序列的任何数据。基于针对分片队列100的多个分片生成的估计访问时间数据450生成有序访问列表452。例如，可以通过基于每个子分片的最早估计访问时间对子分片进行排序来生成有序访问列表452。

[0086] 有序访问列表452可以省略由每个出队会话380-382完全消费的子分片。例如，如图3A-图3B中所示，出队会话380-382已经完全消费了分片102的子分片110-112和分片104的子分片130。因此，从有序访问列表452中省略了子分片110-112和130，这是因为所有出队会话380-382都已经完成了从这些省略的子分片中使消息出队。可替代地和/或附加地，有序访问列表452可以保留已经被所有出队会话380-382完全消费的一个或多个子分片。在这种情况下，水印引用可以被用于指示有序访问列表452中的位置，在该位置之前所有在先的子分片已被所有出队会话380-382完全消费。

[0087] 可以通过一个或多个事件来触发有序访问列表452的重新计算。例如，当特定的出队器-分片对的出队速率显著改变时，可以重新计算有序访问列表452。在这种情况下，重新

计算有序访问列表452可以包括针对特定分片中尚未被特定出队器消费的子分片重新计算一个或多个估计的最早访问时间450。

[0088] 在一些实施例中，当前被访问的子分片(例如，最早估计访问时间是当前时间T_c的子分片)被包括在有序访问列表452中。例如，如由他们的最早访问时间T_c所指示的，子分片114、122、138和152是当前被访问的子分片。出队会话380当前从子分片114进行出队，而出队会话382当前从子分片138进行出队。此外，入队会话202-204当前将消息入队到子分片152中，而其它入队会话(未示出)当前将消息入队到子分片122和138中。

[0089] 在一些实施例中，数据库服务器412实现多个分片队列，并且针对多个分片队列的所有子分片生成统一的有序访问列表，而不考虑子分片所属的分片队列。例如，数据库服务器412的一个或多个后台进程的单个集合可以使用统一的有序访问列表来同时对多个分片队列执行消息高速缓存管理。即，可以基于多个分片队列的子分片的估计访问时间数据来逐出和/或恢复多个分片队列的特定子分片，而不考虑该特定子分片所属的分片队列。

[0090] 跳过列表实施方案

[0091] 在一些实施例中，消息高速缓存414和有序访问列表452被实现为易失性存储器410中的随机化跳过列表。如本文所使用的，术语“跳过列表”是指允许在元素的有序序列内快速搜索的数据结构。跳过列表的每个节点表示来自任何分片102-106的子分片。跳过列表的节点按对应子分片的最早估计访问时间进行排序。跳过列表的头部处节点表示当前被访问的子分片和/或即将被访问的子分片。靠近跳过列表的尾部的节点表示在较远的将来将被访问的子分片。当修改有序访问列表452的次序时，随机化跳过列表中的节点(例如，子分片)之间的链接被修改以反映新的次序。

[0092] 逐出经高速缓存的子分片

[0093] 当确定经高速缓存的子分片不可能被很快访问时，执行逐出经高速缓存的子分片。在一些实施例中，仅响应于确定需要释放消息高速缓存414和/或易失性存储器410中的空间而执行逐出。例如，可以在消息高速缓存414使用的存储器的量达到阈值(例如，可用存储器的50％)时执行逐出。

[0094] 在一些实施例中，在逐出范围内可能不被访问的经高速缓存的子分片是逐出的候选。即，如果估计特定的经高速缓存的子分片在由逐出范围指示的时间帧内不会被访问，那么可以从消息高速缓存414中逐出该特定的子分片。在一些实施例中，逐出范围指示从当前时间开始的时间帧。在这个时间帧期间可能将被访问的子分片被维护在消息队列中。逐出范围可以是缺省值，和/或由管理员或另一个操作者动态设置的值。

[0095] 诸如由于用于将子分片数据写入数据库结构的辅助存储装置I/O操作和/或数据库命令的执行，经高速缓存的子分片的逐出可能花费大量时间和/或其它计算资源。由于这些原因，逐出经高速缓存的子分片会在由出队会话执行时造成延迟增加。在一些实施例中，逐出由一个或多个背景进程执行，这将在下文中更详细地描述。除了逐出范围之外，时间帧还可以基于逐出滞后因子。逐出滞后因子调整逐出范围，以为逐出操作给出足够保守的一段时间。

[0096] 逐出范围和/或逐出滞后因子可以是缺省值、自适应值和/或由管理员或另一个操作者动态设置的值。逐出范围和/或逐出滞后因子可以基于逐出统计信息(诸如逐出一个或多个经高速缓存的子分片所花费的时间量)来确定。在一些实施例中，逐出范围和/或逐出

滞后因子是周期性重新计算的动态值。在一些实施例中，逐出范围和/或逐出滞后因子还基于一个或多个系统特点，诸如子分片尺寸、硬件配置、性能、可用资源、队列尺寸和/或其它系统特点。

[0097] 在一些实施例中，当逐出经高速缓存的子分片时，子分片的表示在辅助存储装置中被存储交换表408中。例如，每个子分片表示可以作为行存储在交换表408中。交换表408包括行，每行包括被逐出的子分片116、134和136的表示。在一些实施例中，存储在辅助存储装置中的子分片的表示是对应的经高速缓存的子分片的二进制表示。在一些实施例中，消息高速缓存414中的子分片存储块被直接复制到辅助存储装置402中，诸如复制到交换表408的行中。消息高速缓存可以具有一个或多个交换表408。在一些实施例中，消息高速缓存存储来自一个或多个不同队列的消息。在这种情况下，这一个或多个不同队列可以共享一个或多个交换表408，或者可以实现特定于该一个或多个不同队列的子集的一个或多个私有交换表408。

[0098] 将子分片恢复到消息高速缓存

[0099] 当恢复先前被逐出的子分片时，所存储的表示被用于在消息高速缓存414中生成经高速缓存的子分片。例如，可以通过将子分片的二进制从交换表408读取到易失性存储器410中来恢复交换表408中的存储的子分片的二进制表示。在一些实施例中，将辅助存储装置402中的子分片存储块直接复制到消息高速缓存414中。

[0100] 可以响应于访问子分片的实际请求而恢复先前被逐出的该子分片。当响应于实际请求而不是在请求之前恢复子分片时，必须首先完成I/O操作以便从辅助存储装置检索子分片。因此，期望使得响应于实际请求而恢复子分片的发生最小化，诸如通过在实际请求之前预获取被逐出的子分片。

[0101] 如本文所使用的，术语“预获取”是指在访问子分片的任何实际请求(诸如执行涉及该子分片的队列操作的请求或指令)之前恢复先前被逐出的该子分片。如果消息高速缓存中有空间并且很有可能在不久的将来会对未经高速缓存的子分片执行队列操作(诸如出队操作)，那么可以预获取该子分片。在一些实施例中，当确定可能在时间帧内访问先前被逐出的子分片时，预获取该子分片。

[0102] 由于诸如用于从数据库结构检索子分片数据的辅助存储装置I/O操作和/或数据库命令的执行，恢复经高速缓存的子分片可能花费大量时间。因为恢复经高速缓存的子分片可能需要花费大量时间，所以在被出队会话执行时它可能造成延迟增加。在一些实施例中，预获取是由一个或多个后台进程来执行的，这将在下文中更详细地描述。

[0103] 在一些实施例中，用于预获取子分片的时间帧基于由预获取范围指示的时间帧。预获取范围指示相对于当前时间的时间帧。在一些实施例中，在这个时间帧期间可能被访问的子分片被恢复到消息队列。预获取范围可以与逐出范围相同或不同。

[0104] 预获取范围的长度是折中。如果预获取范围太短，那么出队会话会遇到在请求出队操作之前未被恢复的子分片。如果预获取范围太长，那么消息高速缓存会变得被当前未访问的经高速缓存的子分片过度填充。除了预获取范围之外，时间帧还可以基于预获取滞后因子。预获取滞后因子调整预获取范围，以便为预获取操作给出足够保守的一段时间。

[0105] 预获取范围和/或预获取滞后因子可以是缺省值、自适应值和/或由管理员或另一个操作者动态设置的值。预获取滞后因子和/或预获取范围可以基于恢复统计信息，诸如恢

复和/或预获取一个或多个经高速缓存的子分片所花费的时间量。在一些实施例中，预获取滞后因子和/或预获取范围是周期性重新计算的动态值。在一些实施例中，预获取滞后因子和/或预获取范围还基于一个或多个系统特点，诸如子分片尺寸、硬件配置、性能、可用资源、队列尺寸和/或其它系统特点。

[0106] 无序消息

[0107] 在一些实施例中，分片队列100被配置为处置无序消息。无序消息包括指示除按照入队时间之外的递送时间的递送指示。无序消息的示例包括具有指示的延迟时间或安排的递送时间的消息。

[0108] 在一些实施例中，基于入队时间将无序消息存储在子分片中。当基于入队时间将无序消息入队时，不能仅基于出队速率和当前出队位置来确定一些估计访问时间数据(诸如最早估计访问时间)。此外，在一些实施例中，在完全消费当前子分片之前，出队会话可能需要从另一个子分片进行出队。

[0109] 在一些实施例中，包含无序消息的特定子分片的最早估计访问时间还基于对应分片中一个或多个无序消息的递送指示。在一些实施例中，最早估计访问时间基于每个出队器已经出队的特定子分片的分数(fraction)。在这种情况下，由每个出队会话处理的特定子分片的分数被维护，因为出队会话可能潜在地无序地进行出队。

[0110] 后台监视进程

[0111] 在一些实施例中，一个或多个后台监视进程监视出队会话以检测出队器是否已经在子分片当中拖延(stall)。在处理子分片的过程中，出队会话有时会减慢，完全拖延和/或失败。在这种情况下(诸如在出队器在子分片之间过渡时进行更新的实施例中)，可以不更新对应的出队器-分片出队速率。同样，出队会话可以在子分片内加速。为了防止这种改变以影响消息队列管理的方式影响估计访问时间数据，后台监视进程可以监视子分片内的出队会话进度。

[0112] 例如，后台监视进程可以确定是否有任何出队器在处理子分片时已经减慢、拖延、失败、加速或以其它方式改变速度。当后台监视进程确定出队器在处理特定子分片时已经改变速度时，后台监视进程可以更新对应的出队器-分片出队速率、估计访问时间数据450和/或其它统计信息和/或引发这些更新。

[0113] 后台监视进程可以通过以预定义次序和/或自适应次序检查出队会话418来监视出队进度。在一些实施例中，后台监视进程基于最早估计访问时间450从可能是下一个被访问的子分片的子分片开始遍历有序访问列表452。对于尚未完全消费子分片的每个出队会话，后台监视进程将该出队会话的当前位置与相应的出队器-分片出队速率进行比较，并计算自上次更新以来出队会话应当已经遍历的子分片的数量。如果出队会话应当已经遍历了比出队会话的当前位置所指示的更多子分片或更少子分片，那么出队会话可能已经减慢、拖延、失败、加速或以其它方式改变了速度。在一些实施例中，当检测到这种状况时，更新对应的出队器-分片对的出队速率，并且重新计算有序访问列表452和/或其它估计访问时间数据450。

[0114] 后台逐出进程

[0115] 可以由一个或多个后台逐出进程来执行逐出操作。后台逐出进程可以(诸如基于逐出范围和/或存储器考虑因素)识别要逐出的经高速缓存的子分片。在一些实施例中，仅

当存在存储器状况时后台逐出进程才逐出经高速缓存的子分片。

[0116] 在一些实施例中，后台逐出进程使得基于估计访问时间数据可能在最大量的时间后才被访问的经高速缓存的子分片的逐出优先化。例如，后台逐出进程可以从尾部遍历有序访问列表452。后台逐出进程试图逐出所遇到的每个经高速缓存的子分片，直到它到达基于估计访问时间数据可能在由逐出范围和/或逐出滞后因子指示的时间帧内被访问的子分片。在一些实施例中，逐出滞后因子基于后台逐出进程逐出子分片的速率。

[0117] 后台预获取进程

[0118] 可以由一个或多个后台预获取进程来执行预获取操作。因为预获取是在访问特定子分片的任何实际请求之前在该特定子分片上执行的，所以后台预获取进程可以潜在地在供预获取的候选子分片上迭代，而不影响在用于执行涉及子分片的队列操作的实际请求或指令时的性能。

[0119] 在一些实施例中，后台预获取进程使得基于估计访问时间数据将可能在最少量的时间内被访问的被逐出的子分片的预获取优先化。例如，后台预获取进程可以从头部遍历有序访问列表452。后台预获取进程试图预获取每个被逐出的子分片，直到它到达基于估计访问时间数据可能在由预获取范围和/或预获取滞后因子指示的时间帧之后才被访问的子分片。在一些实施例中，预获取滞后因子基于后台预获取进程恢复被逐出的子分片的速率。

[0120] 共享盘数据库实例

[0121] 在一些实施例中，数据库服务器412是分布式数据库系统(诸如共享盘数据库系统)中的实例。在这种情况下，包括辅助存储装置402的存储系统由多个实例共享。持久队列表406在共享存储系统中维护。每个实例在其相应的易失性存储器410中维护本地消息高速缓存414。

[0122] 在一些实施例中，为了使得多实例数据库中的ping最小化，队列分片上的所有消息入队都在单个实例上完成。此外，可以在单个实例上由同一个出队器完成队列分片上的所有消息出队。如果队列分片的入队实例和出队实例不相同，那么交叉进程可以将消息从入队实例发送到出队实例。因此，相同的子分片可以存在于多个实例中。

[0123] 就消息高速缓存管理而言，每个交叉进程可以被建模为入队实例上的出队进程和出队实例中的入队进程。交叉进程的入队速率是交叉进程的出队速率乘以发送到出队实例的消息的百分比。虽然系统不具有对从入队会话提供的入队速率以及从出队会话提供的出队速率的控制，但系统可以控制每个交叉进程的短期出队速率(以及因此控制短期入队速率)以帮助管理共享盘数据库系统中消息高速缓存的短期存储器需求。

[0124] 示例处理

[0125] 图6是图示根据一个或多个实施例的、用于消息高速缓存管理的示例处理的流程图。处理600可以由一个或多个计算设备和/或其进程执行。例如，处理600的一个或多个框可以由计算设备800执行。在一些实施例中，处理600的一个或多个框由数据库系统(诸如数据库系统400)执行。

[0126] 在框602处，系统将来自多个入队器的多个消息入队在分片队列中。在一些实施例中，每个入队器具有与分片队列的特定分片的入队亲和性，并且由入队器入队的消息在该特定分片的当前子分片中入队在该特定分片的尾部处。

[0127] 在框604处，系统维护被配置为存储多个经高速缓存的子分片的消息高速缓存。经

高速缓存的子分片与分片队列的存储在存储器中的一组子分片对应以便于队列操作。

[0128] 在框606处，系统针对多个出队器-分片对中的每个出队器-分片对确定出队器对于该分片的出队速率。在一些实施例中，出队速率基于指定的出队器在一段时间内已处理的指定分片的子分片的数量。

[0129] 在框608处，系统基于出队器-分片对的出队速率针对多个子分片中的每个子分片生成估计访问时间数据。在一些实施例中，子分片的估计访问时间数据是由分片队列的任何出队器对于该子分片的最早估计访问时间。

[0130] 在框610处，系统基于出队器-分片对出队速率信息确定要作为经高速缓存的子分片存储在消息高速缓存中的一组子分片。例如，系统可以基于根据出队器-分片对的出队速率生成的估计访问时间数据来确定要高速缓存哪些子分片。在一些实施例中，系统基于多个子分片中未被所有出队器完全消费的每个子分片的最早估计访问时间来确定哪些子分片要存储在消息高速缓存中。该一组子分片可以基于消息高速缓存和/或系统中的可用存储器的量。选择经高速缓存的子分片以最小化逐出操作和恢复操作，同时确保当任何出队器执行出队操作时经高速缓存的子分片的可用性。

[0131] 在框612处，处理600返回和/或终止。例如，处理可以通过将控制传递给调用进程、生成任何适当的记录或通知、在方法或函数调用之后返回、或者终止来继续。

[0132] 图7是图示根据一个或多个实施例的、用于消息高速缓存管理的处理的实施例的流程图。处理700可以由一个或多个计算设备和/或其进程执行。例如，处理700的一个或多个框可以由计算设备800执行。在一些实施例中，处理700的一个或多个框由数据库系统(诸如数据库系统400)执行。

[0133] 在框702处，系统基于出队器-分片速率信息和/或从出队器-分片对生成的基于出队器-分片速率信息生成的估计访问时间数据，来确定要从消息高速缓存中逐出的一个或多个经高速缓存的子分片。

[0134] 在框704处，系统将经高速缓存的子分片的表示存储在辅助存储装置中。在一些实施例中，系统使用数据库命令将经高速缓存的子分片的表示存储在辅助存储装置中的数据库结构中。例如，系统可以将经高速缓存的子分片的表示存储为数据库的交换表中的行。

[0135] 在框706处，系统从消息高速缓存中移除经高速缓存的子分片。

[0136] 在框708处，系统确定将一个或多个子分片恢复到消息高速缓存。例如，在预期有来自出队器的未来出队请求的情况下，系统可以基于更新后的出队器-分片信息确定应当通过预获取子分片并在存储器高速缓存中生成经高速缓存的子分片来将先前被逐出的子分片恢复在消息高速缓存中。在一些情况下，恢复子分片的确定是响应于来自出队器的实际出队请求而做出的。

[0137] 在框710处，将一个或多个子分片恢复到消息高速缓存。在一些实施例中，通过从辅助存储装置检索子分片的表示来生成经高速缓存的子分片。例如，数据库命令可以被用于从辅助存储装置中的数据库结构(诸如交换表)获取子分片的表示。

[0138] 在框712处，处理700返回和/或终止。例如，处理可以通过将控制传递给调用进程、生成任何适当的记录或通知、在方法或函数调用之后返回、或者终止来继续。

[0139] 数据库系统

[0140] 由于本文描述的一些实施例是在数据库管理系统(DBMS)的上下文中实现的，因此

本文包括对数据库管理系统的描述。DBMS管理数据库。DBMS可以包括一个或多个数据库服务器。数据库包括存储在持久存储机制(诸如硬盘的集合)上的数据库数据和数据库字典。数据库数据可以存储在一个或多个数据容器中,每个数据容器包含一个或多个记录。每个记录内的数据被组织成一个或多个字段。在关系DBMS中,数据容器被称为表,记录被称为行,并且字段被称为列。在面向对象的数据库中,数据容器被称为对象类,记录被称为对象(在本文也被称为对象记录),并且字段被称为属性。其它数据库体系结构可以使用其它术语。

[0141] 用户通过向数据库服务器提交使得数据库服务器对存储在数据库中的数据执行操作的命令来与DBMS的数据库服务器交互。用户可以是在客户端上运行的与数据库服务器交互的一个或多个应用。

[0142] 数据库命令可以是符合数据库语言的语法的数据库语句的形式。用于表达数据库命令的一种示例语言是结构化查询语言(SQL)。SQL数据定义语言(“DDL”)指令被发布到DBMS以定义数据库结构,诸如表、视图或复杂数据类型。例如,CREATE、ALTER、DROP和RENAME是一些SQL实施方案中发现的DDL指令的常见示例。SQL数据操纵语言(“DML”)指令被发布到DBMS以管理存储在数据库结构内的数据。例如,SELECT、INSERT、UPDATE和DELETE是在一些SQL实施方案中找到的DML指令的常见示例。SQL/XML是在对象-关系数据库中操纵XML数据时使用的SQL的常见扩展。

[0143] 在数据库服务器内执行操作常常使得需要调用多层软件。层是软件模块的集合,其执行数据库服务器内的在某种程度上专用于该软件模块集合的功能。执行操作通常涉及调用多层软件,其中一个层调用另一个层,在执行第一调用期间调用另一个层。例如,为了执行SQL语句,将调用SQL层。通常,客户端通过接口(诸如到SQL层的SQL接口)访问数据库服务器。SQL层分析和解析并且执行语句。在执行语句期间,SQL层调用较低层的模块,以从表中检索特定行以及更新表中的特定行。客户端(诸如复制客户端)通常经由对数据库服务器的(诸如以SQL语句的形式的)数据库命令来访问数据库。

[0144] 虽然上述示例基于Oracle的SQL,但是本文提供的技术不限于Oracle的SQL、任何专有形式的SQL、任何标准化版本或形式的SQL(ANSI标准),或任何特定形式的数据库命令或数据库语言。此外,为了简化本文包含的解释,数据库命令或其它形式的计算机指令可以被描述为执行动作,诸如创建表、修改数据和设置会话参数。但是,应当理解的是,数据库命令本身不执行任何操作,而是DBMS在执行数据库命令时执行对应的操作。通常,数据库命令通过与数据库的同步连接来执行。

[0145] 示例实施系统

[0146] 根据一些实施例,本文所描述的技术由一个或多个专用计算设备实现。专用计算设备可以是硬连线的以执行所述技术,或者可以包括诸如被永久性地编程以执行所述技术的一个或多个专用集成电路(ASIC)或现场可编程门阵列(FPGA)的数字电子设备,或者可以包括编程为按照固件、存储器、其它存储装置或者其组合中的程序指令执行所述技术的一个或多个通用硬件处理器。这种专用计算设备还可以合并定制的硬连线逻辑、ASIC或FPGA与定制的编程来实现所述技术。专用计算设备可以是台式计算机系统、便携式计算机系统、手持式设备、联网设备或者结合硬连线和/或程序逻辑来实现所述技术的任何其它设备。

[0147] 例如,图8是描绘可以在其上实现实施例的计算机系统800的框图。计算机系统800

包括总线802或用于传递信息的其它通信机制，以及与总线802耦合以处理信息的硬件处理器804。硬件处理器804可以是例如通用微处理器。

[0148] 计算机系统800还包括耦合到总线802用于存储信息和要由处理器804执行的指令的主存储器806，诸如随机存取存储器(RAM)或其它动态存储设备。主存储器806也可以用于存储在执行由处理器804执行的指令期间的临时变量或其它中间信息。当存储在处理器804可访问的非瞬态存储介质中时，这些指令使计算机系统800成为执行指令中所指定的操作而定制的专用机器。

[0149] 计算机系统800还包括只读存储器(ROM)808或者耦合到总线802的其它静态存储设备，用于为处理器804存储静态信息和指令。提供了存储设备810，诸如磁盘、光盘或固态驱动器，并且将其耦合到总线802，用于存储信息和指令。

[0150] 计算机系统800可以经总线802耦合到显示器812，诸如阴极射线管(CRT)，用于向计算机用户显示信息。包括字母数字和其它键的输入设备814耦合到总线802，用于向处理器804传送信息和命令选择。另一种类型的用户输入设备是游标控件816，诸如鼠标、轨迹球或者游标方向键，用于向处理器804传送方向信息和命令选择并且用于控制显示器812上的游标运动。这种输入设备通常具有在两个轴(第一个轴(例如，x)和第二个轴(例如，y))中的两个自由度，这允许设备在平面内指定位置。

[0151] 计算机系统800可以使用定制的硬连线逻辑、一个或多个ASIC或FPGA、固件和/或程序逻辑来实现本文所述的技术，这些与计算机系统相结合，使计算机系统800或者把计算机系统800编程为专用机器。根据一个实施例，本文的技术由计算机系统800响应于处理器804执行包含在主存储器806中的一条或多条指令的一个或多个序列而执行。这些指令可以从另一个存储介质(诸如存储设备810)读到主存储器806中。包含在主存储器806中的指令序列的执行使处理器804执行本文所述的处理步骤。在替代实施例中，硬连线的电路系统可以代替软件指令或者与其结合使用。

[0152] 如在本文所使用的，术语“存储介质”是指存储使机器以特定方式操作的数据和/或指令的任何非瞬态介质。这种存储介质可以包括非易失性介质和/或易失性介质。非易失性介质包括例如光盘、磁盘或固体驱动器，诸如存储设备810。易失性介质包括动态存储器，诸如主存储器806。存储介质的常见形式包括，例如，软盘、柔性盘、硬盘、固态驱动器、磁带，或者任何其它磁性数据存储介质，CD-ROM，任何其它光学数据存储介质，任何具有孔模式的物理介质，RAM、PROM和EPROM、FLASH-EPROM、NVRAM，任何其它存储器芯片或盒带。

[0153] 存储介质与传输介质不同但是可以与其结合使用。传输介质参与在存储介质之间传送信息。例如，传输介质包括同轴线缆、铜线和光纤，包括包含总线802的配线。传输介质还可以采取声波或光波的形式，诸如在无线电波和红外线数据通信中产生的那些声波或光波。

[0154] 各种形式的介质可以参与把一条或多条指令的一个或多个序列携带到处理器804供执行。例如，指令最初可以在远程计算机的磁盘或固态驱动器上携带。远程计算机可以把指令加载到其动态存储器中并且利用调制解调器经电话线发送指令。位于计算机系统800本地的调制解调器可以接收电话线上的数据并且使用红外线发送器把数据转换成红外线信号。红外线检测器可以接收在红外线信号中携带的数据并且适当的电路系统可以把数据放在总线802上。总线802把数据携带到主存储器806，处理器804从该主存储器806检索并执

行指令。由主存储器806接收的指令可以可选地在被处理器804执行之前或之后存储在存储设备810上。

[0155] 计算机系统800还包括耦合到总线802的通信接口818。通信接口818提供耦合到网络链路820的双向数据通信，其中网络链路820连接到本地网络822。例如，通信接口818可以是综合业务数字网络(ISDN)卡、电缆调制解调器、卫星调制解调器，或者提供到对应类型的电话线的数据通信连接的调制解调器。作为另一个示例，通信接口818可以是提供到兼容的局域网(LAN)的数据通信连接的LAN卡。也可以实现无线链路。在任何此类实施方案中，通信接口818都发送和接收携带表示各种类型信息的数字信号流的电、电磁或光信号。

[0156] 网络链路820通常通过一个或多个网络向其它数据设备提供数据通信。例如。网络链路820可以通过本地网络822提供到主计算机824或者到由互联网服务提供商(ISP)826操作的数据装备的连接。ISP 826又通过现在通常称为“互联网”828的全局分组数据通信网络提供数据通信服务。本地网络822和互联网828二者都使用携带数字数据流的电、电磁或光信号。通过各种网络的信号以及在网络链路820上并通过通信接口818的信号是传输介质的示例形式，其中信号把数字数据带到计算机系统800或者携带来自计算机系统800的数字数据。

[0157] 计算机系统800可以通过(一个或多个)网络、网络链路820和通信接口818发送消息和接收数据，包括程序代码。在互联网示例中，服务器830可以通过互联网828、ISP 826、本地网络822和通信接口818发送对应于应用程序的所请求代码。

[0158] 接收到的代码可以在其被接收到时由处理器804执行，和/或存储在存储设备810或其它非易失性存储装置中，用于以后执行。

[0159] 在前面的说明书中，本发明的实施例已经参考众多的具体细节进行了描述，这些细节可以因实施方案而异。因此，说明书和附图被认为是说明性的而不是限制性的。本发明的范围的唯一且排他指示，以及申请人所预期的作为本发明的范围的内容，是从本申请产生的权利要求集合的书面和等效范围，以这种权利要求产生的具体形式，包括任何后续的更正。

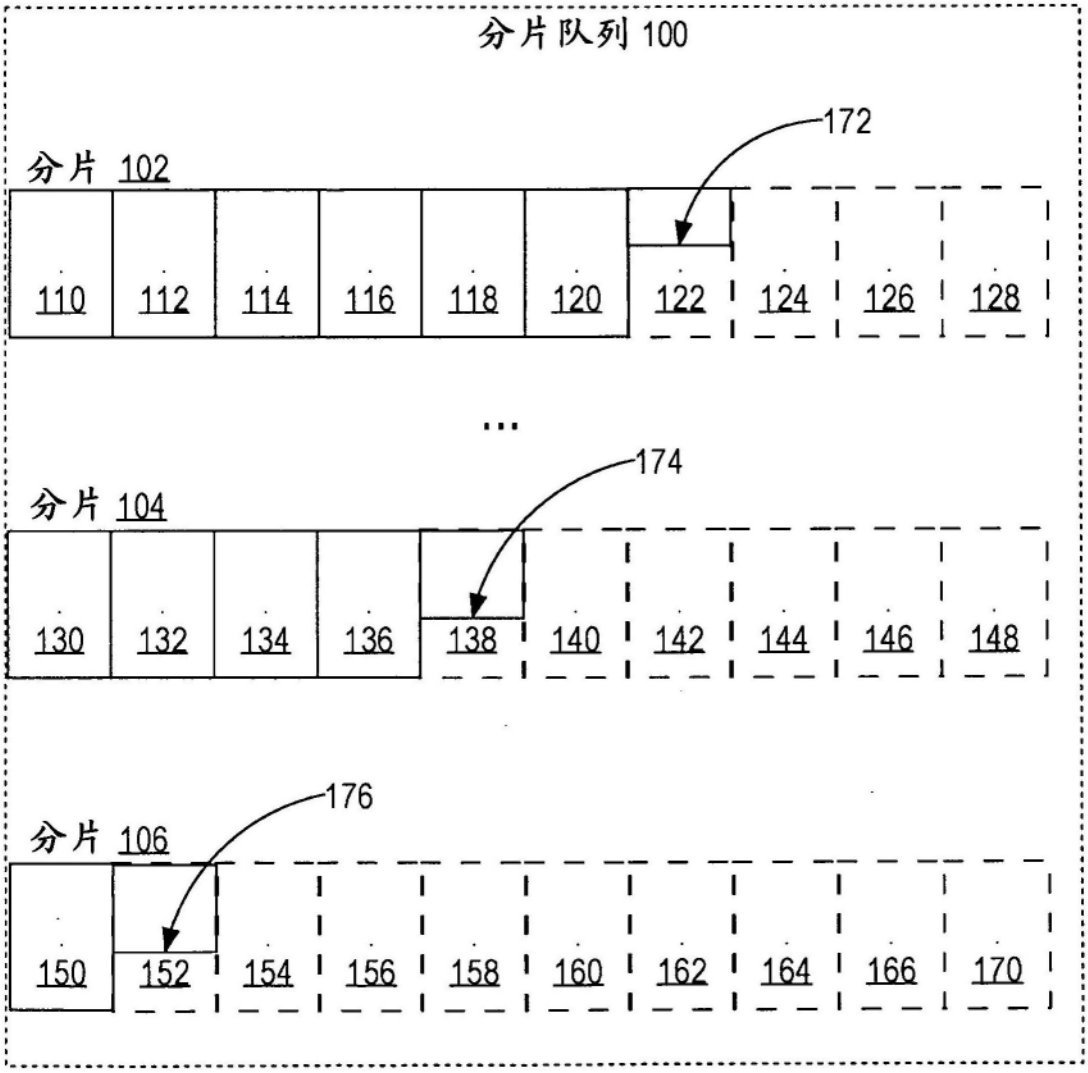

分片队列 100
分片 102
172
110
112
114
116
118
120
122
124
126
128
...
分片 104
174
130
132
134
136
138
140
142
144
146
148
分片 106
176
150
152
154
156
158
160
162
164
166
170

图1

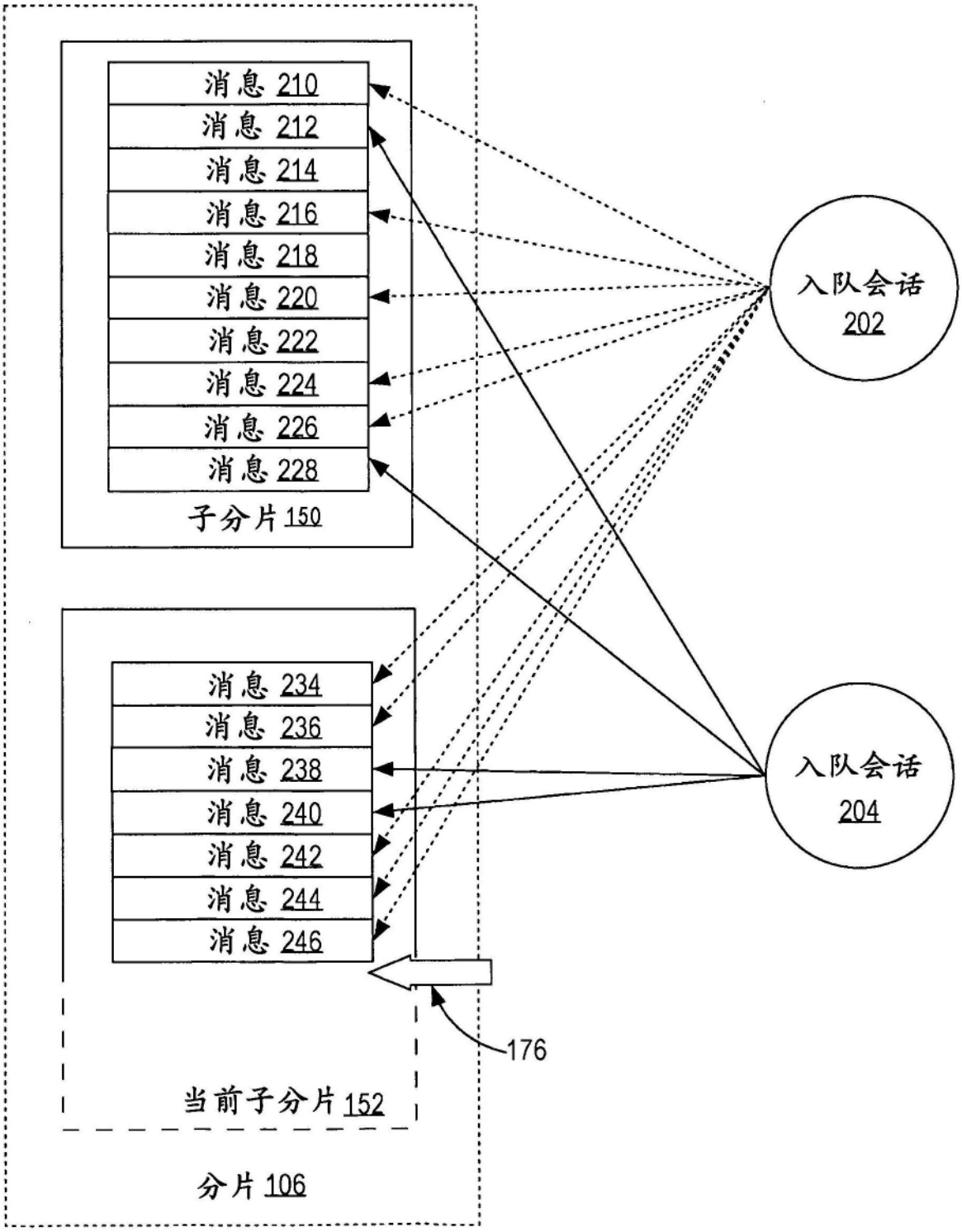

消息 210
消息 212
消息 214
消息 216
消息 218
消息 220
消息 222
消息 224
消息 226
消息 228
子分片 150
消息 234
消息 236
消息 238
消息 240
消息 242
消息 244
消息 246
当前子分片 152
分片 106
176
入队会话 202
入队会话 204

图2

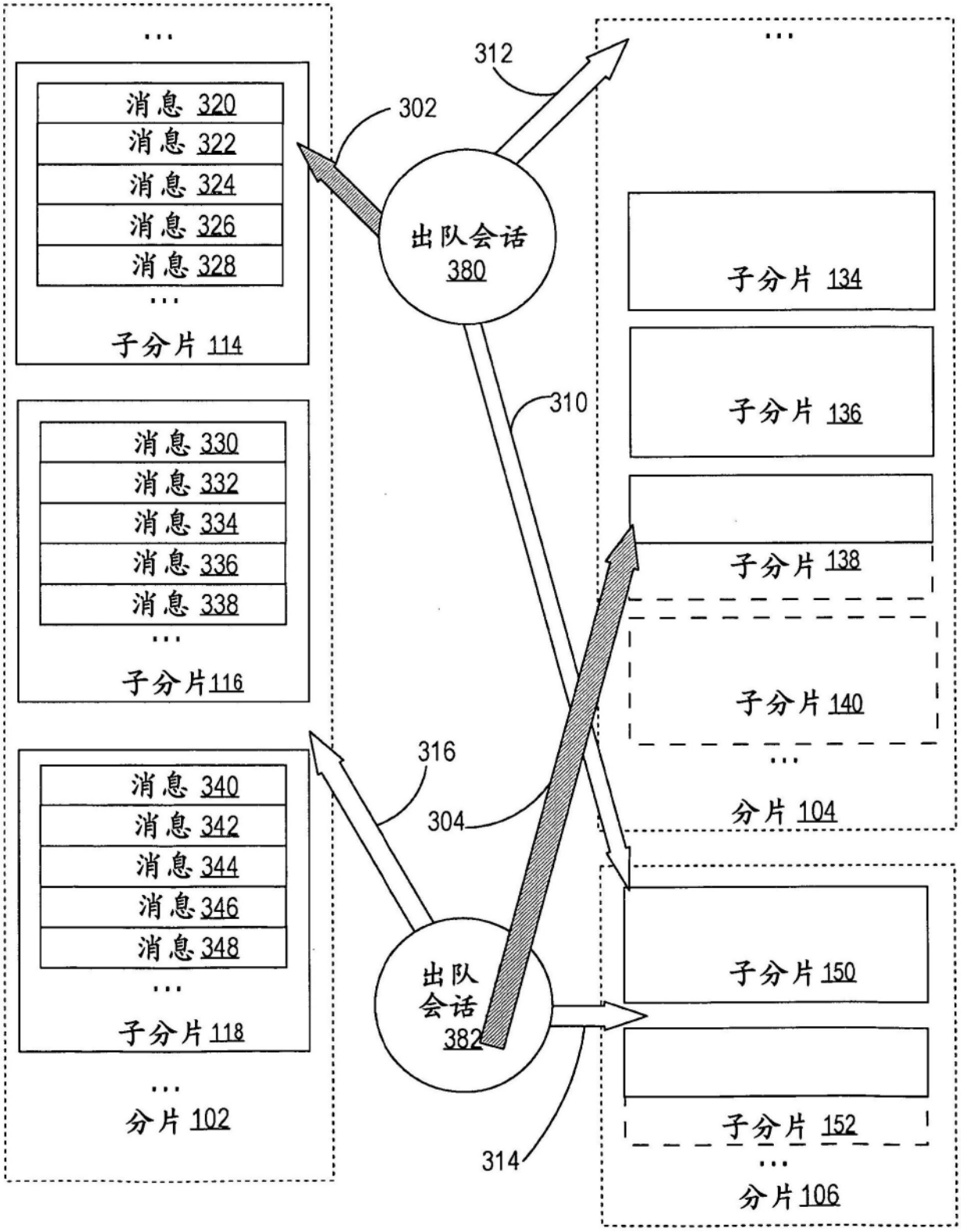

消息 320
消息 322
消息 324
消息 326
消息 328
子分片 114
消息 330
消息 332
消息 334
消息 336
消息 338
子分片116
消息 340
消息 342
消息 344
消息346
消息348
子分片 118
分片 102
302
312
出队会话 380
310
子分片 134
子分片 136
子分片 138
子分片 140
分片 104
316
304
出队会话 382
子分片 150
子分片 152
分片106
314

图3A

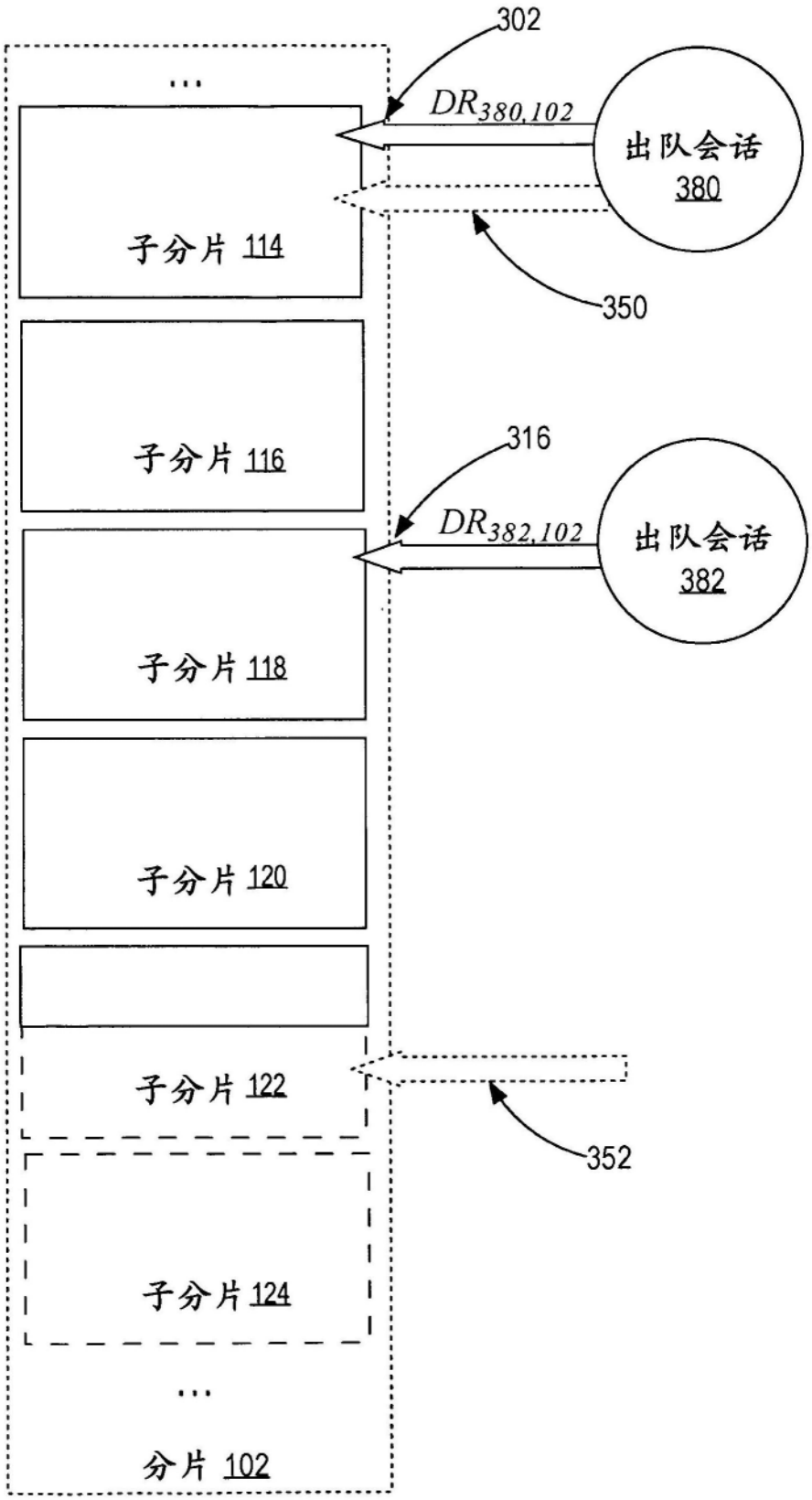

302
$DR_{380,102}$
出队会话
380
350
子分片 114
子分片 116
316
$DR_{382,102}$
出队会话
382
子分片 118
子分片 120
子分片 122
352
子分片 124
分片 102

图3B

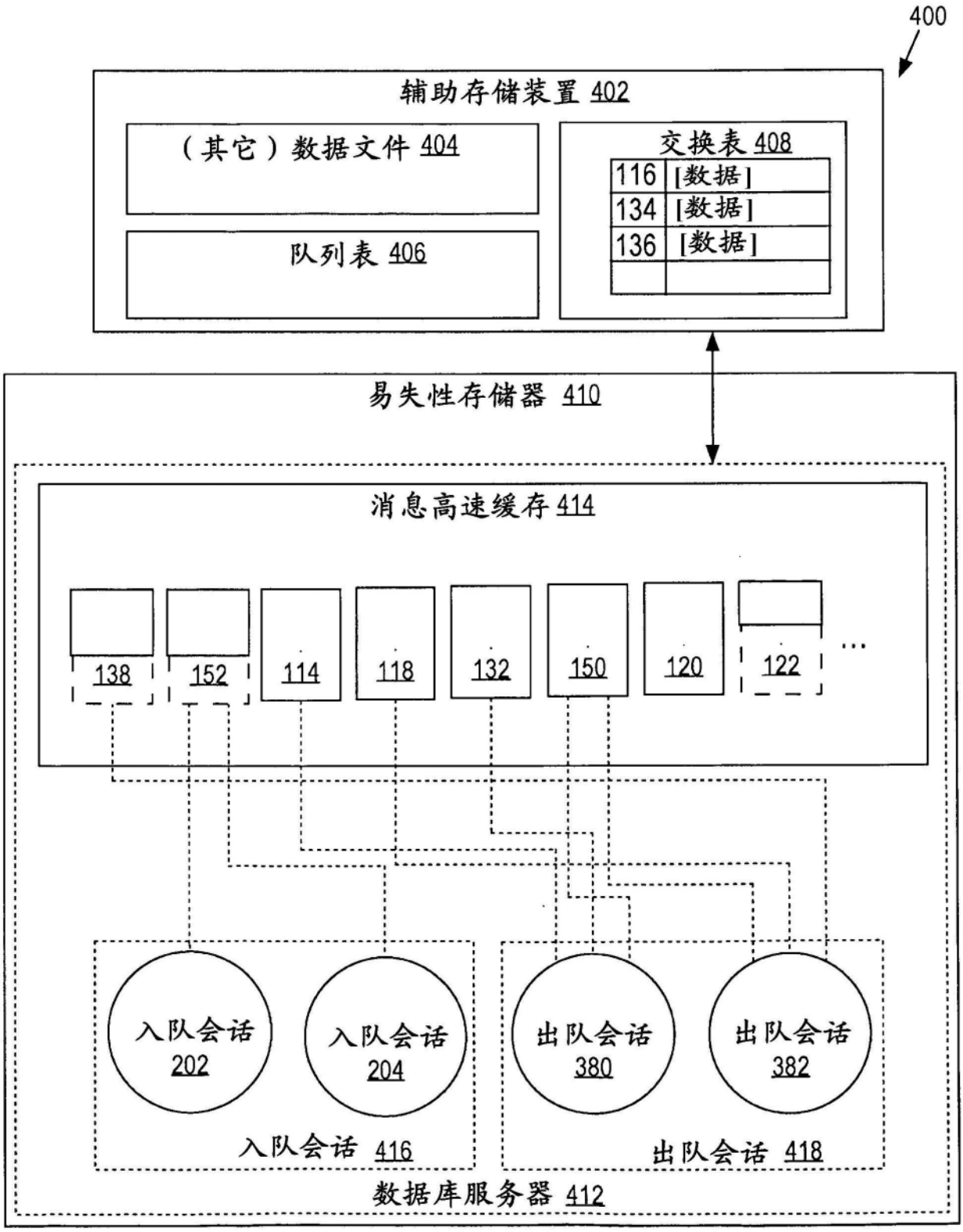

400
辅助存储装置 402
(其它)数据文件 404
队列表 406
交换表 408
116 [数据]
134 [数据]
136 [数据]
易失性存储器 410
消息高速缓存 414
138
152
114
118
132
150
120
122
入队会话 202
入队会话 204
出队会话 380
出队会话 382
入队会话 416
出队会话 418
数据库服务器 412

图4A

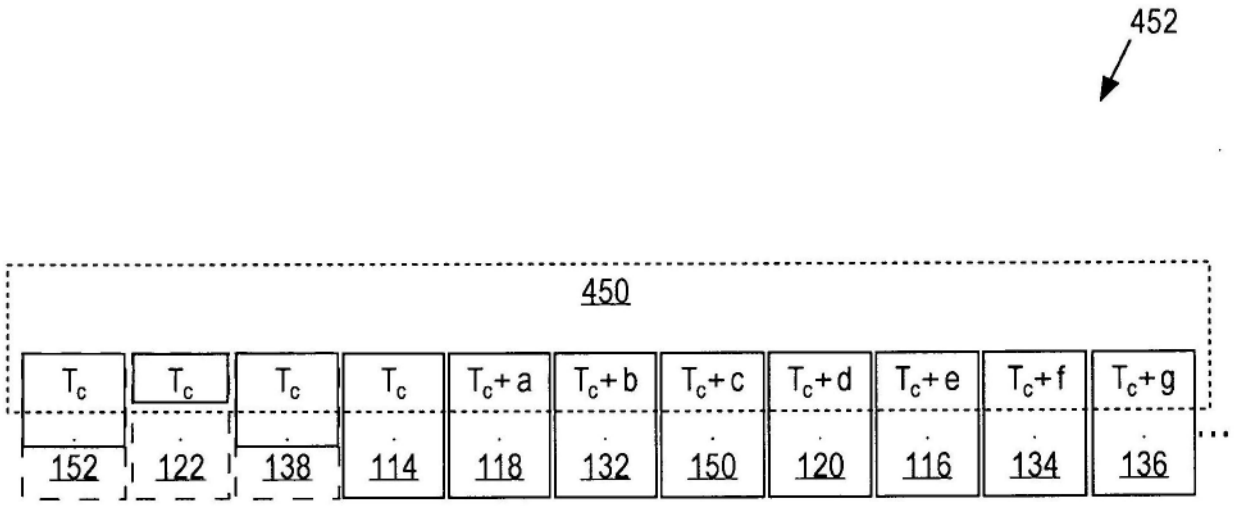

452
450
T_c
T_c
T_c
T_c
T_c+a
T_c+b
T_c+c
T_c+d
T_c+e
T_c+f
T_c+g
152
122
138
114
118
132
150
120
116
134
136

图4B

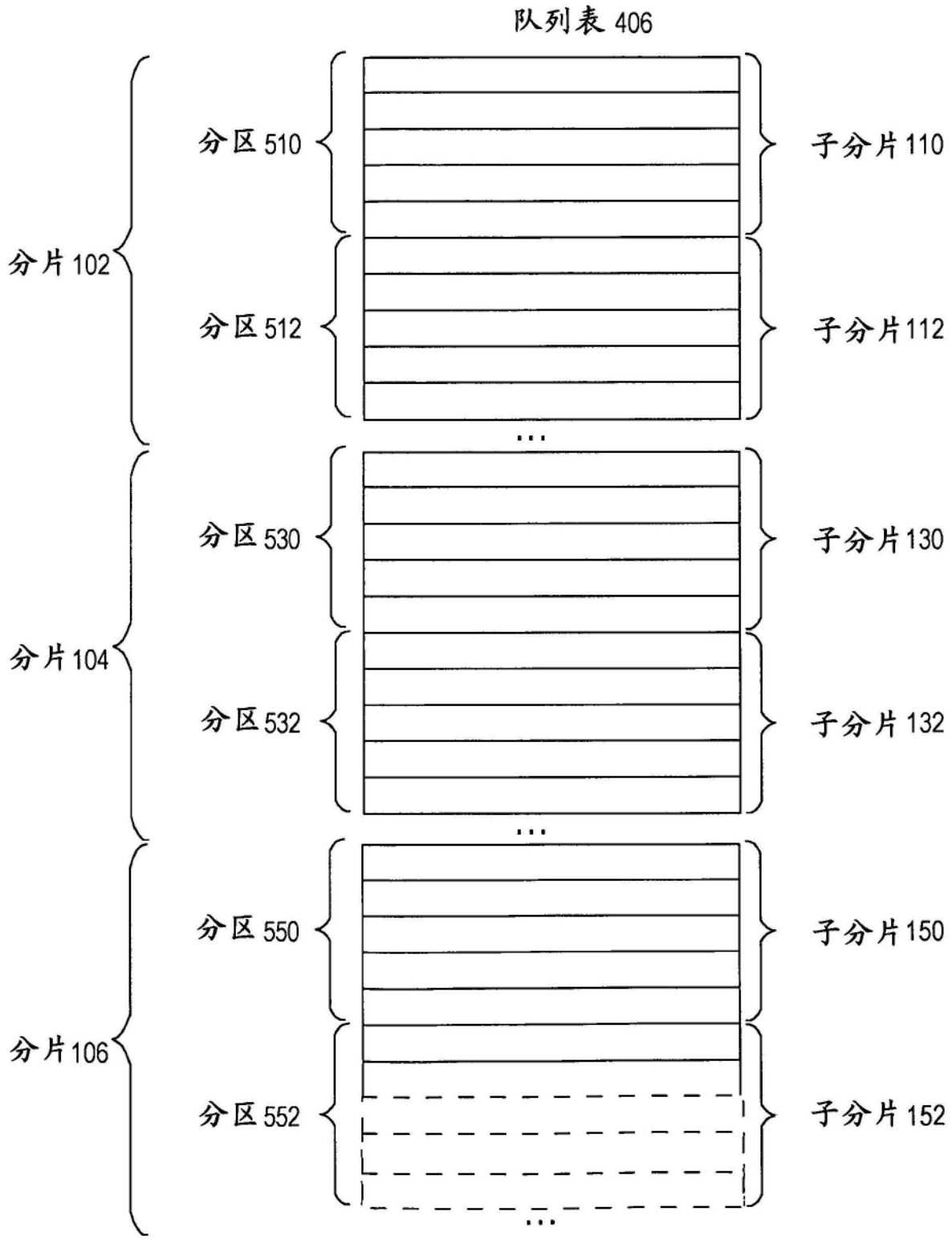

队列表 406
分片 102
分区 510
子分片 110
分区 512
子分片 112
分片 104
分区 530
子分片 130
分区 532
子分片 132
分片 106
分区 550
子分片 150
分区 552
子分片 152

图5

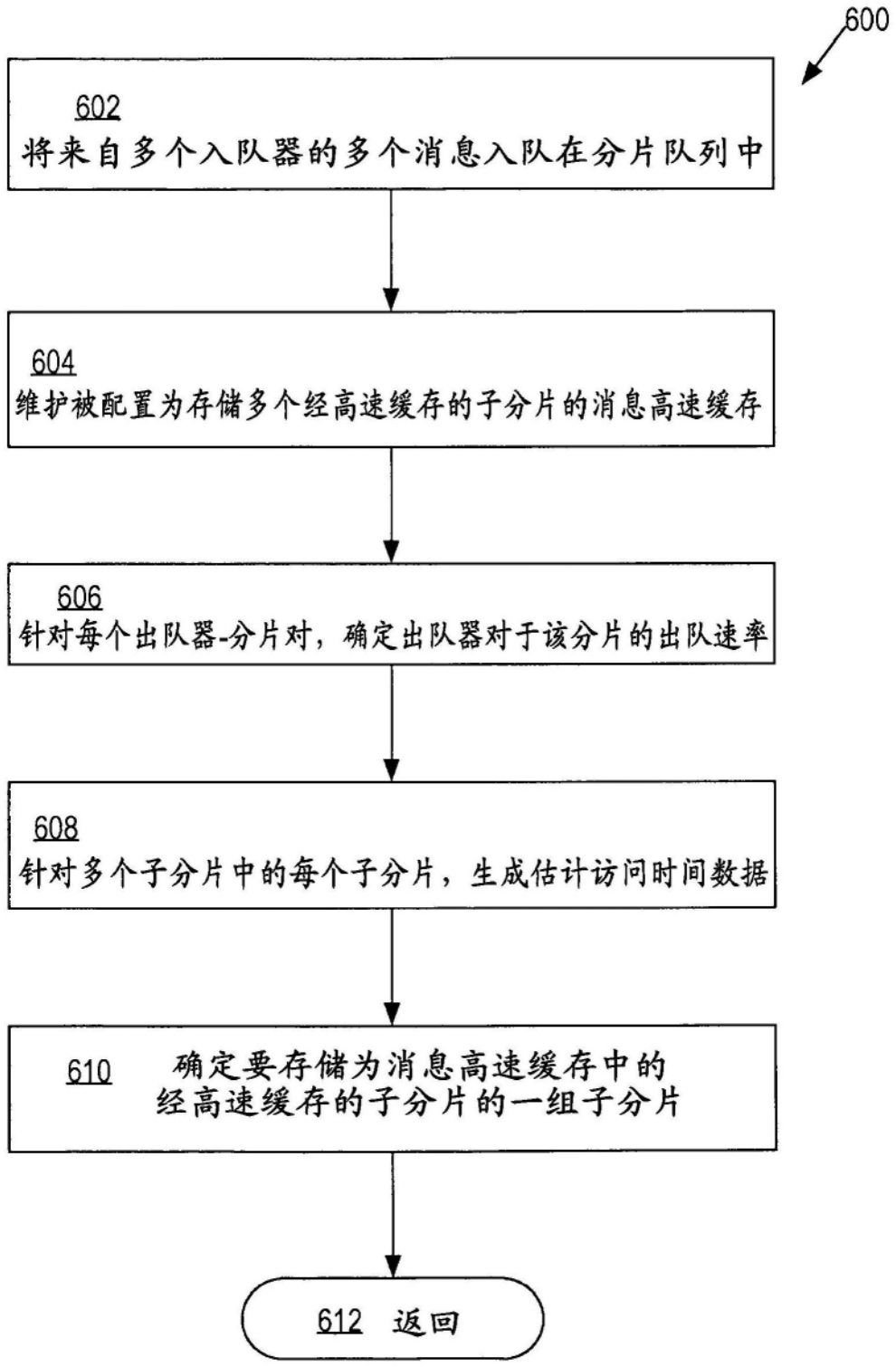

600
602
将来自多个入队器的多个消息入队在分片队列中
604
维护被配置为存储多个经高速缓存的子分片的消息高速缓存
606
针对每个出队器-分片对，确定出队器对于该分片的出队速率
608
针对多个子分片中的每个子分片，生成估计访问时间数据
610
确定要存储为消息高速缓存中的经高速缓存的子分片的一组子分片
612 返回

图6

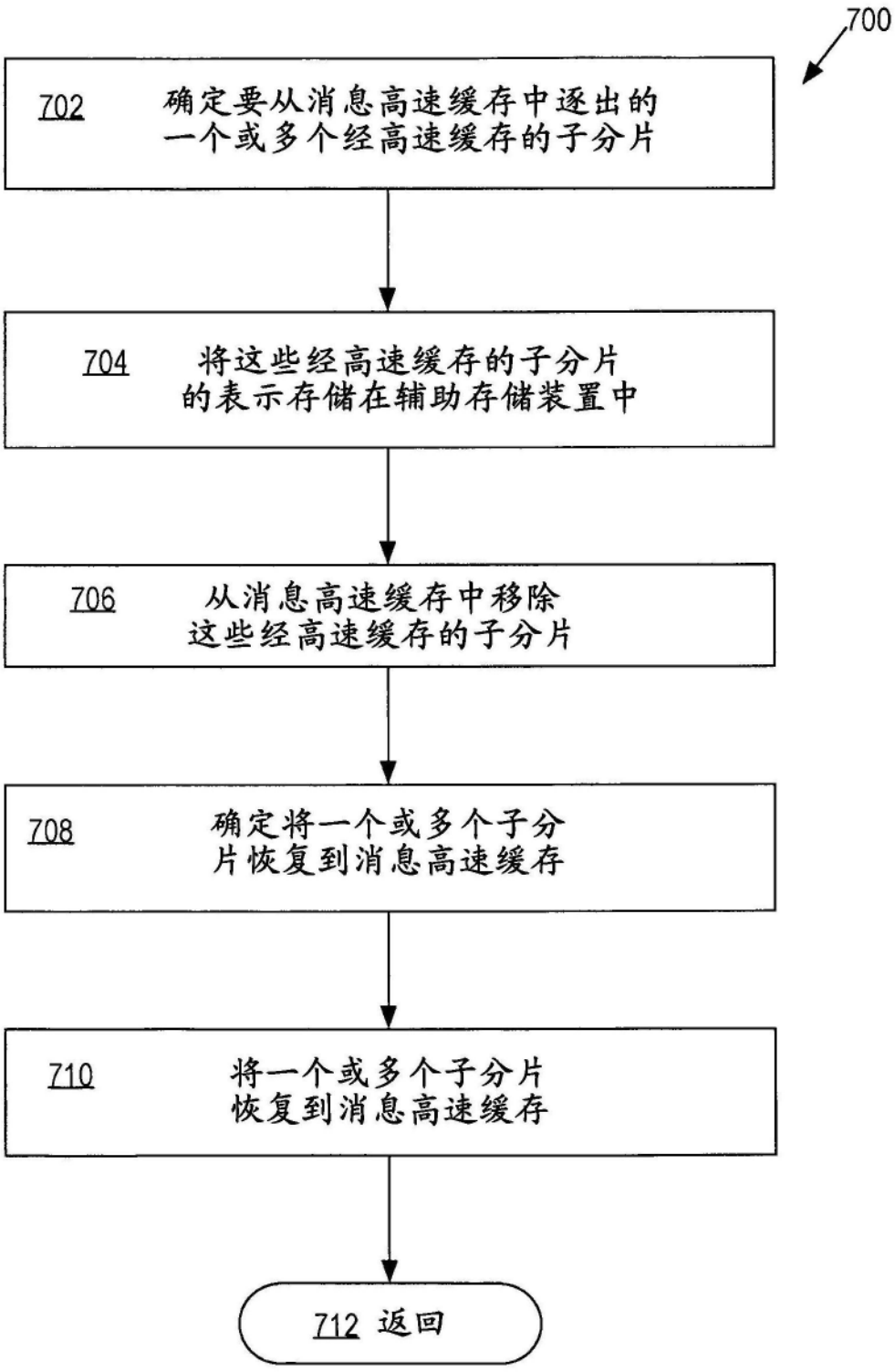

700
702
确定要从消息高速缓存中逐出的一个或多个经高速缓存的子分片
704
将这些经高速缓存的子分片的表示存储在辅助存储装置中
706
从消息高速缓存中移除这些经高速缓存的子分片
708
确定将一个或多个子分片恢复到消息高速缓存
710
将一个或多个子分片恢复到消息高速缓存
712 返回

图7

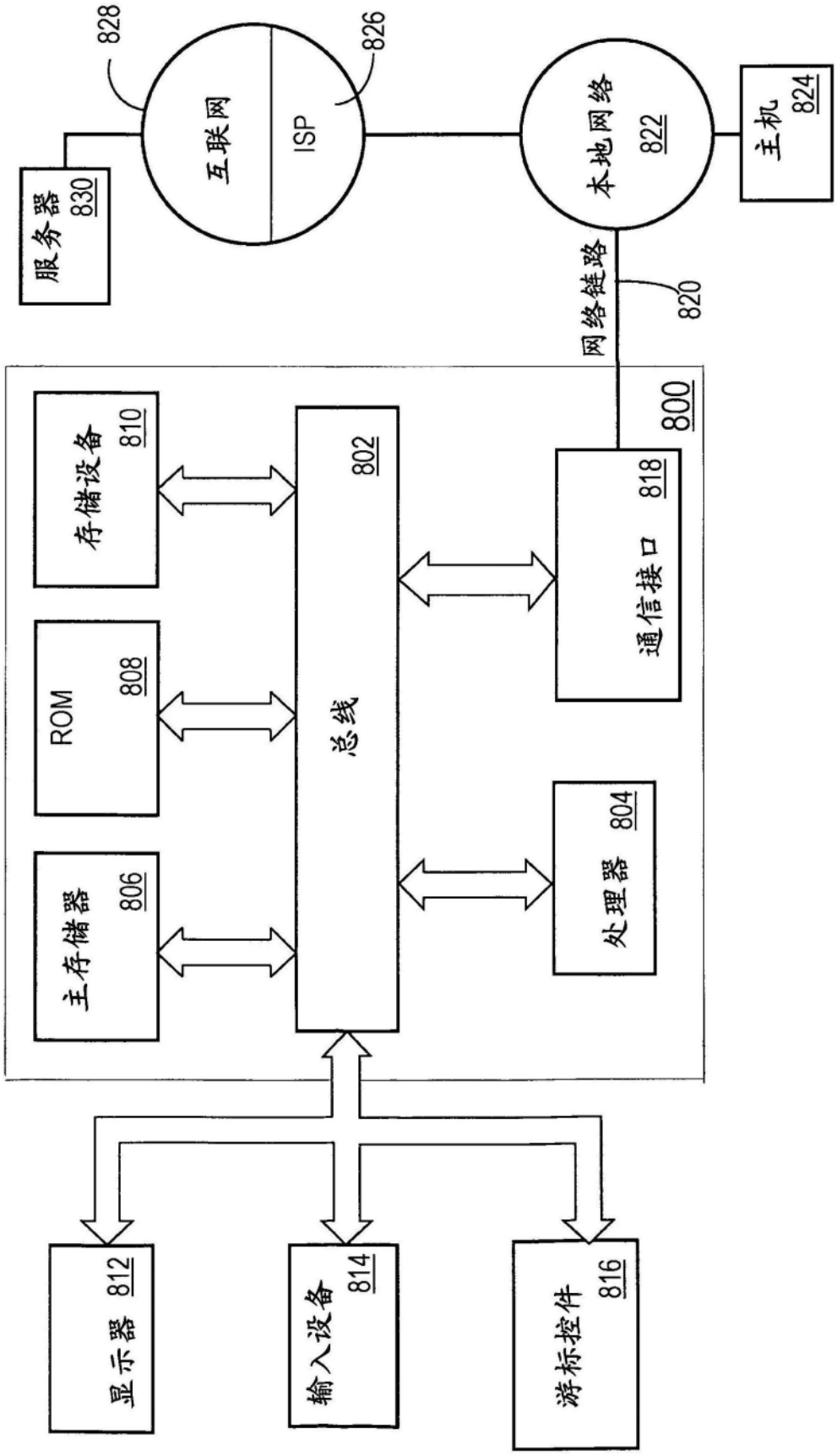

828
互联网
ISP
826
本地网络
822
主机
824
服务器
830
网络链路
820
800
存储设备
810
802
通信接口
818
ROM
808
总线
主存储器
806
处理器
804
显示器
812
输入设备
814
游标控件
816

图8