

[19] 中华人民共和国国家知识产权局



[12] 发明专利申请公布说明书

[21] 申请号 200480042560.8

[51] Int. Cl.

G06F 11/00 (2006.01)

G06F 11/07 (2006.01)

G06F 11/20 (2006.01)

[43] 公开日 2007 年 11 月 21 日

[11] 公开号 CN 101076784A

[22] 申请日 2004.12.10

[21] 申请号 200480042560.8

[30] 优先权

[32] 2004.3.25 [33] US [31] 10/808,839

[86] 国际申请 PCT/US2004/041240 2004.12.10

[87] 国际公布 WO2005/101991 英 2005.11.3

[85] 进入国家阶段日期 2006.9.25

[71] 申请人 伊姆西公司

地址 美国马萨诸塞州

[72] 发明人 斯蒂芬·斯特里克兰

约翰·V··伯勒斯 蒂莫西·多尔

[74] 专利代理机构 北京金信立方知识产权代理有限公司

代理人 黄威

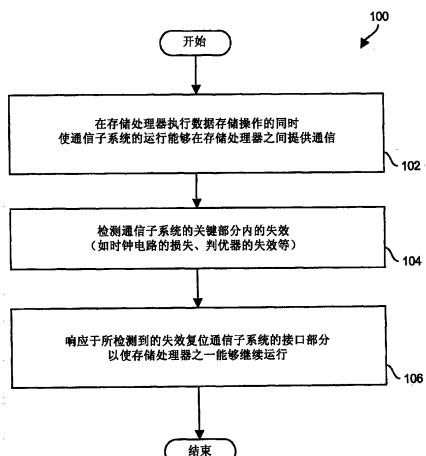
权利要求书 3 页 说明书 8 页 附图 4 页

[54] 发明名称

在失效期间维持数据存储系统运行的技术

[57] 摘要

数据存储系统具有第一存储处理器、第二存储处理器和通信子系统。通信子系统具有(i)互连在第一存储处理器和第二存储处理器之间的接口部分，(ii)连到接口部分的时钟电路，及(iii)连到接口部分和时钟电路的控制器。控制器被配置成使接口部分的运行能够在第一和第二存储处理器之间提供通信、检测时钟电路内的失效、并响应于所检测的失效复位接口部分以使第一和第二存储处理器之一能够继续运行。接口部分的这种复位防止其余存储处理器锁止，因而释放该存储处理器，从而即使在失效之后也能继续运行。



1、数据存储系统，包括：

第一存储处理器；

第二存储处理器；和

通信子系统，其具有（i）互连在第一存储处理器和第二存储处理器之间的接口部分，（ii）连到接口部分的时钟电路，及（iii）连到接口部分和时钟电路的控制器；控制器被配置成：

使接口部分的运行能够在第一和第二存储处理器之间提供通信；

检测时钟电路内的失效；及

响应于所检测到的失效复位接口部分以使第一和第二存储处理器之一能够继续运行。

2、根据权利要求 1 的数据存储系统，其中通信子系统的控制器包括：

把关器阶段，其被配置成响应于预定超时周期内时钟电路的时钟信号损失而产生错误信号。

3、根据权利要求 2 的数据存储系统，其中通信子系统的接口部分包括连到第一存储处理器的第一接口设备、连到第二存储处理器的第二接口设备、及将第一和第二接口设备连在一起的通信总线；且其中通信子系统的控制器还包括：

连到把关器阶段的输出阶段，输出阶段被配置成响应于错误信号而向第一接口设备提供复位信号，复位信号使第二存储处理器能够继续运行。

4、根据权利要求 1 的数据存储系统，其中通信子系统的接口部分包括：

高速缓存镜像接口（CMI）总线；

第一接口设备，其具有连到第一存储处理器的第一 PCI 接口和连到 CMI 总线的第一 CMI 接口；及

第二接口设备，其具有连到第二存储处理器的第二 PCI 接口和连到 CMI 总线的第二 CMI 接口。

5、根据权利要求 1 的数据存储系统，其中通信子系统的接口部分包括：

连到第一存储处理器的第一接口；

连到第二存储处理器的第二接口；及

连到通信子系统的控制器的开关，开关位于第一和第二接口之间。

6、根据权利要求 5 的数据存储系统，其中第一存储处理器从第一电源接收功率，其中第二存储处理器从第二电源接收功率，且其中通信子系统的控制器还被配置成：

响应于第一和第二电源之一的电源信号损失而断开开关。

7、具有第一存储处理器和第二存储处理器的数据存储系统的通信子系统，该通信子系统包括：

被配置以使第一存储处理器和第二存储处理器互连的接口部分；

连到接口部分的时钟电路；及

连到接口部分和时钟电路的控制器，控制器被配置成：

使接口部分的运行能够在第一和第二存储处理器之间提供通信；

检测时钟电路内的失效；及

响应于所检测到的失效复位接口部分以使第一和第二存储处理器之一能够继续运行。

8、在具有(i)第一存储处理器、(ii)第二存储处理器、及(iii)连到第一和第二存储处理器的通信子系统的数据存储系统中，用于在通信子系统内出现失效期间运行数据存储系统的方法，该方法包括：

在第一和第二存储处理器执行数据存储操作的同时，使通信子系统的运行能够在第一和第二存储处理器之间提供通信；

检测通信子系统的关键部分内的失效；及

响应于所检测到的失效复位通信子系统的接口部分以使第一和第二存储处理器之一能够继续运行。

9、根据权利要求 8 的方法，其中通信子系统的关键部分包括时钟电路；其中检测失效包括：

响应于预定超时周期内时钟电路的时钟信号损失而产生错误信号；

其中通信子系统包括连到第一存储处理器的第一接口设备、及连到第二存储处理器的第二接口设备，第一和第二接口设备通过通信总线连接在一起；且其中复位接口部分包括：

向第一接口设备输出复位信号以使第二存储处理器能够继续运行。

10、根据权利要求 8 的方法，其中通信子系统的接口部分包括连到第一存储处理器的第一接口和连到第二存储处理器的第二接口；其中该方法还包括：

响应于所检测到的失效断开第一和第二接口之间的开关；

其中通信子系统的关键部分包括 (i) 被配置为从第一存储处理器的第一电源接收第一电源信号的第一电源输入，及 (ii) 被配置为从第二存储处理器的第二电源接收第二电源信号的第二电源输入；且其中断开开关包括：

响应于第一和第二电源信号之一的损失而中断第一和第二接口之间的电学通路。

在失效期间维持数据存储系统运行的技术

背景

数据存储系统代表一个或多个外部主计算机保存和检索信息。典型的数据存储系统包括网络适配器、存储处理电路、一组磁盘驱动器。网络适配器在外部主计算机和存储处理电路之间提供连通性。存储处理电路执行多种数据存储操作（如装入操作、保存操作、读-修改-写操作等），并提供高速缓冲存储器，高速缓冲存储器使数据存储系统能够优化其操作（如提供高速保存、数据预取等）。磁盘驱动器组提供坚固的数据存储容量，但其是以较慢且非易失的方式提供。

某些数据存储系统的存储处理电路包括多个存储处理单元，以实现更大的可用性和/或更大的数据存储吞吐量。在这样的系统中，每一存储处理单元均能独立执行数据存储操作。

例如，一个常规的数据存储系统包括两个存储处理单元，其被配置成通过高速缓存镜像接口（CMI）相互通信以维持高速缓存一致并使高速缓存镜像磁盘写的影响最小。具体地，CMI 总线使一份数据在磁盘写操作完成之前可为两个存储处理单元所用。在该系统中，第一存储处理单元具有第一 CMI 接口电路，第二存储处理单元具有第二 CMI 接口电路，第一和第二 CMI 接口电路通过 CMI 总线相互连接。

发明内容

不幸地，对上述常规数据存储系统有某些限制。例如，在该数据存储系统运行期间，在有关 CMI 的电路内可能有失效（如时钟失效、判优器失效等）或数据处理单元之一中有失效。例如，假如 CMI 接口电路之一正在 CMI 总线上发出指令的过程中，而这时相对的 CMI 接口电路中出现失效。在这种情况下，使 CMI 接口电路不失效的机会中止，继而锁止其存储处理单元的运行。如果其发生，则整个数据存储系统将被阻止执行进一步的数据存储操作。

另外，大多数具有多个存储处理器的常规数据存储系统包括花费昂贵的、具有多个电源的冗余电源装备，使得，如果一个电源失效，则该失效将不会使系统不工作。不幸地，如果用相对廉价的标准电源代替该冗余电源装备，则存在这样的风险：用户不注意而拔出 AC 线并导致不是电源失效的功率损失，因而损害其它没有失效的电路（如存储处理器）。

相比于上述常规数据存储系统，本发明的实施例致力于在失效（如布置在存储处理器之间的一部分通信子系统内的单点失效）期间维持具有多个存储处理器的数据存储系统的运行的技术。具体地，这样的技术预防不注意地锁止其余存储处理器以保持整个数据存储系统的可用性（即使存储处理器能够继续运行）。另外，这样的技术使能够使用相对便宜的标准电源向每一存储处理器单独供电，并为共享资源如通信子系统局部提供共享功率，从而既节约成本又具有可靠的失效容错度。也就是说，这些技术使能够使用低成本的商品零件以降低总成本，而不会危及整体可靠性。

本发明的一实施例为具有第一存储处理器、第二存储处理器和通信子系统的数据存储系统。通信子系统具有 (i) 互连在第一存储处理器和第二存储处理器之间的接口部分，(ii) 连到接口部分的时钟电路，及 (iii) 连到接口部分和时钟电路的控制器。控制器被配置成使接口部分的运行能够在第一和第二存储处理器之间提供通信、检测时钟电路内的失效、并响应于所检测的失效复位接口部分以使第一和第二存储处理器之一能够继续运行。接口部分的这种复位防止其余存储处理器锁止，因而释放该存储处理器，从而即使在失效之后也能继续运行。

在一种方案中，通信子系统的接口部分包括连到第一存储处理器的第一接口、连到第二存储处理器的第二接口、及连到通信子系统的控制器的开关。开关位于第一和第二接口之间。在该方案中，控制器被配置成响应于或来自供电第一接口的第一电源或来自供电第二接

口的第二电源的电源损失信号而断开开关。因而，其余接口提供的任何电压将不会损害已失去功率的接口。

附图说明

本发明的前述及其它目标、特征和优点将从下面结合附图给出的本发明特定实施例的描述中看出，其中在不同的图中同一附图标记指同一组件。附图不必定按比例绘制，而是强调本发明原理的图示。

图 1 为适于本发明使用的数据存储系统的方块图。

图 2 为图 1 的数据存储系统的通信子系统的一部分的方块图。

图 3 为图 1 的数据存储系统的通信子系统的另一部分的方块图。

图 4 为通信子系统在失效期间所执行的过程的流程图。

具体实施方式

本发明的实施例致力于在失效（如布置在存储处理器之间的一部分通信子系统内的单点失效）期间维持具有多个存储处理器的数据存储系统的运行的技术。具体地，这样的技术预防不注意地锁止其余存储处理器以保持整个数据存储系统的可用性（即，使存储处理器能够继续运行）。另外，这样的技术使能够使用相对便宜的标准电源向每一存储处理器单独供电，并为共享资源如通信子系统局部提供共享功率，从而既节约成本又具有可靠的失效容错度。也就是说，这些技术使能够使用低成本的商品零件以降低总成本，而不会危及整体可靠性。

图 1 所示为适于本发明使用的数据存储系统 20。数据存储系统 20 被配置成代表一组外部主机 22 (1)、…、22 (n)（统称主机 22）保存和检索信息。数据存储系统 20 包括一个或多个网络接口（为简化起见未示出），以使数据存储系统 20 能够使用各种不同的协议与主机 22 通信，这些协议如：TCP/IP 通信协议、光纤通道协议、计数-键码-数据（CKD）记录格式协议、I/O 块协议等。

如图 1 中所示，数据存储系统 20 包括处理电路 24 和一批存储装置 26（如磁盘驱动器）。处理电路 24 包括存储处理器 28 (A)、28 (B)（统称为存储处理器 28）和位于存储处理器 28 之间的高速缓存镜像接口（CMI）通信子系统 30。存储处理器 28 被配置成代表主机 22 一个一个单独执行数据存储操作。存储处理器 28 被配置成通过 CMI 通信子系统 30 相互通信。具体地，存储处理器 28 根据 CMI 协议交换指令和数据以维持高速缓存相关性并使高速缓存镜像对整个系统性能的影响最小。

进一步地，如图 1 中所示，存储处理器 28 (A) 包括电源 32 (A)、局部时钟 34 (A)、控制电路 36 (A)、及另外的逻辑电路 38 (A)。控制电路 36 (A) 实质上是存储处理器 28 (A) 的处理引擎，因为其基于来自电源 32 (A) 的电源信号 40 (A) 和来自局部时钟 34 (A) 的时钟信号 42 (A) 执行数据存储操作（如装入和保存操作、高速缓存操作等）。应理解的是，为了简化，将这些信号 40 (A)、42 (A) 传输给控制电路 36 (A) 的特定电源层/线路和时钟迹线在图 1 中已被故意省略。

类似地，存储处理器 28 (B) 包括电源 32 (B)、局部时钟 34 (B)、控制电路 36 (B)、及另外的逻辑电路 38 (B)。连同存储处理器 28 (B) 一起，控制电路 36 (B)（即处理引擎）由来自电源 32 (B) 的电源信号 40 (B) 供电并由来自局部时钟 34 (B) 的时钟信号 42 (B) 驱动。再次声明，为了简化，将这些信号 40 (B)、42 (B) 传输给控制电路 36 (B) 的特定电源层/线路和时钟迹线在图 1 中已被故意省略。

如图 1 进一步所图示的，通信子系统 30 包括共用功率源 44、接口部分 46 和控制部分 48。共用功率源 44 从电源 32 (A)、32 (B)（统称为电源 32）接收功率信号 40 (A)、40 (B)（统称为功率信号 40），并将共用功率（即局部共享的功率）提供给通信子系统 30 的各个组成部分。因此，如果电源 32 之一发生失效，在其余电源 32 提供的功率的基础上，各个组成部分应能够继续运行。

接口部分 46 互连在存储处理器 28 (A) 和存储处理器 28 (B) 之间，并在存储处理器 26 之间提供 CMI 通信通路，以使存储处理器 26 能够协调其运行。控制部分 48 控制接口部分 46 的运行。下面还将更详细地对通信子系统 30 进行阐述。

接口部分 46 包括连到第一存储处理器 28 (A) 的第一接口设备 50 (A)、连到第二存储处理器 28 (B) 的第二接口设备 50 (B)、及将接口设备 50 (A)、50 (B) (统称为接口设备 50) 连在一起的 CMI 总线 52。举例来说，每一接口设备 50 是密封的通用件，其一侧提供 CMI 接口，另一侧提供 PCI 接口。因此，控制电路 36 (A)、36 (B) (统称为控制电路 36) 通过总线 54 连到接口设备 50，总线 54 为局部 PCI 总线。

为支撑接口设备 50 的运行，通信子系统 30 的控制部分 48 包括时钟电路 56、控制器 58、把关器电路 60 和开关 62。时钟电路 56 被配置成输出共用时钟信号 64。与时钟电路 56 连接的接口设备 50 使用共用时钟信号 64 用于通过 CMI 总线 52 进行的通信，并使用局部时钟信号 42 (A)、42 (B) (统称为局部时钟信号 42) 用于通过局部总线 54 进行的通信。通过接口设备 50 的虚线意于图示说明接口设备 50 基于这些时钟信号 64、42 的局部同步运行。

与时钟电路 56 和接口设备 50 连接的控制器 58 被配置成使能接口部分 46 (即接口设备 50) 的运行，因而使能通过 CMI 总线 52 在存储处理器 28 之间进行通信。为防止通信子系统 30 锁止整个数据存储系统 20，控制器 58 被配置成检测和处理某些临界特性的失效。例如，控制器 58 被配置成检测时钟电路 56 内的失效 (如时钟信号 64 的损失)，并响应于所检测到的失效复位接口部分 46 以使存储处理器 28 之一能够继续运行，从而维持数据存储系统 20 的整体可用性。关于该特征的详细情况还将结合图 2 进行说明。

图 2 所示为通信子系统 30 的控制器 58 和把关器电路 60。控制器 58 包括时钟输入 70、判优器电路 72 和除法器 74。把关器电路 60 包括把关器阶段 76 和输出阶段 78。把关器阶段 76 包括分别对应于

存储处理器 28 (A)、28 (B) 的各个把关器元件 80 (A)、80 (B) (统称为把关器元件 80)。类似地，输出阶段 78 包括各个输出元件 82 (A)、82 (B) (统称为输出元件 82)，其分别连到接口设备 50 (A)、50 (B) 因而分别对应于存储处理器 28 (A)、28 (B)。

在运行期间，时钟输入 70 从时钟电路 56 接收共用时钟信号 64，判优器电路 72 根据 CMI 协议协调存储处理器 28 之间的操作。另外，除法器 74 (如计数器) 计数时钟信号 64 的时钟脉冲，并分别将除法器信号 84 (A)、84 (B) (统称为除法器信号 84) 输出给把关器元件 80。每一除法器信号 84 具有较时钟信号 64 长的周期。在一种方案中，除法器 74 是除以 32 电路，其将时钟频率截除为 32 段。在其它方案中，除法器 74 为除以 64 电路，其将时钟频率截除为 64 段。

把关器阶段 76 的把关器元件 80 监视除法器信号 84 以监视心跳即时钟脉冲，如果在预定时间周期内 (如几秒) 没有看见时钟脉冲，则遵照接口设备 50 行事。具体地，把关器元件 80 (A) 向输出元件 82 (A) 提供控制信号 86 (A)，其控制输出信号 88 (A) 是使能还是复位存储处理器 28 (A) 的接口设备 50 (A)。类似地，把关器元件 80 (B) 向输出元件 82 (B) 提供控制信号 86 (B)，其控制输出信号 88 (B) 是使能还是复位存储处理器 28 (B) 的接口设备 50 (B)。

该操作使把关器电路 60 能够复位接口部分 46，从而在时钟电路 44 或判优器电路 72 有失效时避免中止整个数据存储系统 20。具体地，只要把关器元件 80 在预定时间周期内接收时间脉冲，把关器元件 80 指示输出元件 82 使能接口设备 50 的运行。然而，如果把关器元件 80 (如输出元件 82 (B)) 因未能在超时周期内接收到时钟脉冲而超时，则把关器元件 80 输出错误信号 (如控制信号 86 (B) 的不同电压)，其使得相应的输出元件 82 (如输出元件 82 (B)) 输出复位信号 (如输出信号 88 (B) 内的复位脉冲，见图 2)，从而复位其各自的接口设备 50 (如接口设备 50 (B))。在一种方案中，接口设备 50 保持复位模式，直到整个数据存储系统 20 执行恢复或复位过程为止。

如上所述，在通信子系统 30 内单点失效之后（如时钟电路 56 或判优器 72 失效），以允许存储处理器 28（如存储处理器 28（B））以容错方式维持运行的方式，复位的接口设备 50 被有效地停用。也就是说，存储处理器 28 不由其接口设备 50 锁止，而是能够继续代表主机 22 执行数据存储操作。本发明的实施例的进一步详细描述将结合图 3 给出。

图 3 示出了控制器 58 的另一部分 90。如图所示，控制器 58 的部分 90 包括电压监控器 92（A）、92（B），其分别连到存储处理器 28（A）、28（B）的电源 32（A）、32（B）以接收电源信号 40（A）、40（B）。电压监控器 92（A）、92（B）（统称为电压监控器 92）还连到沿 CMI 总线 52 布置的开关 62（参见图 1）。

部分 90 被配置成控制 CMI 总线 52 的电学通路的连通性。具体地，只要部分 90 接收到电源信号 40（A）、40（B），部分 90 提供开关信号 94（A）、94（B），其闭合开关 62 因而连接接口 50。

然而，假设电源 32 之一（如电源 32（B））失效。在这种情形下，当相应的电压监控器 92（如电压监控器 92（B））未能接收到其各自的电源信号 40（如电源信号 40（B））时，电压监控器 92 断开开关 62（如改变开关信号 94（B）的电压）以中断 CMI 总线 52 的电学通路。因此，失效的存储处理器 28 的接口设备 50 不会被其余存储处理器 28 的接口设备 50 的电压输出损害（例如，当接口设备 50（B）未被供电时，接口设备 50（B）的输出驱动器不会被接口设备 50（A）提供的电压永久损害）。此外，CMI 总线 52 的拉起将防止接口设备 50（A）免遭继续的损害。由于没有长期损害，与从失效恢复正常相关的时间耗费、努力及成本均将最小。本发明实施例的进一步详细描述将结合图 4 给出。

图 4 是在特定失效期间通信子系统 30 的把关器电路 60 的运行过程 100 的流程图。在步骤 102，当存储处理器 28 执行数据存储操作时，把关器电路 60 使通信子系统 30 的接口设备 50 能够在存储处理器 28 之间提供 CMI 通信。

在步骤 104，把关器电路 60 检测通信子系统的关键部分内的失效。例如，把关器电路 60 确定时钟电路 56 或判优器 72 是否已失效。

在步骤 106，把关器电路 60 响应于所检测到的失效复位通信子系统 30 的接口部分 46，以使存储处理器 28 之一能够继续运行。这样的运行使得即使在发生失效之后数据存储系统 20 仍能保持可用。

如上所述，本发明的实施例致力于在失效（如位于存储处理器 28 之间的一部分通信子系统 30 内的单点失效）期间维持具有多个存储处理器 28 的数据存储系统 20 的运行的技术。具体地，这样的技术预防不注意地锁止其余存储处理器 28 以保持整个数据存储系统 20 的可用性（即，使存储处理器 28 能够继续运行）。另外，这样的技术使能够使用相对便宜的标准电源 32 (A)、32 (B) 向每一存储处理器 28 (A)、28 (B) 单独供电，并为共享资源如通信子系统 30 局部提供共享功率，从而既节约成本又具有可靠的失效容错度。也就是说，这些技术使能够使用低成本的商品零件以降低总成本，而不会危及整体可靠性。

在本发明已结合其优选实施例进行特别示出和描述的同时，本领域技术人员应该理解的是，在不脱离后附权利要求确定的本发明精神和范围的情况下可进行各种形式和细节的变化。

例如，应理解的是，仅作为例子，存储处理电路 24 之间的通信通路在上面被阐释为 CMI 总线。其它通信通路也适于使用，如标准通信通道，包括 PCI 总线、GP/IO 线路、无线通路、光学通路等。

另外，应理解的是，仅作为例子，数据存储系统 20 在上面被描述为包括两个存储处理器 28。在其它方案中，数据存储系统 20 具有不同数量的存储处理器 28（如三个、四个等）。此外，这些方案可包括不同的通信配置，如多点总线协议，而不是 CMI 通道。这样的修改和增强均属于本发明的不同实施例。

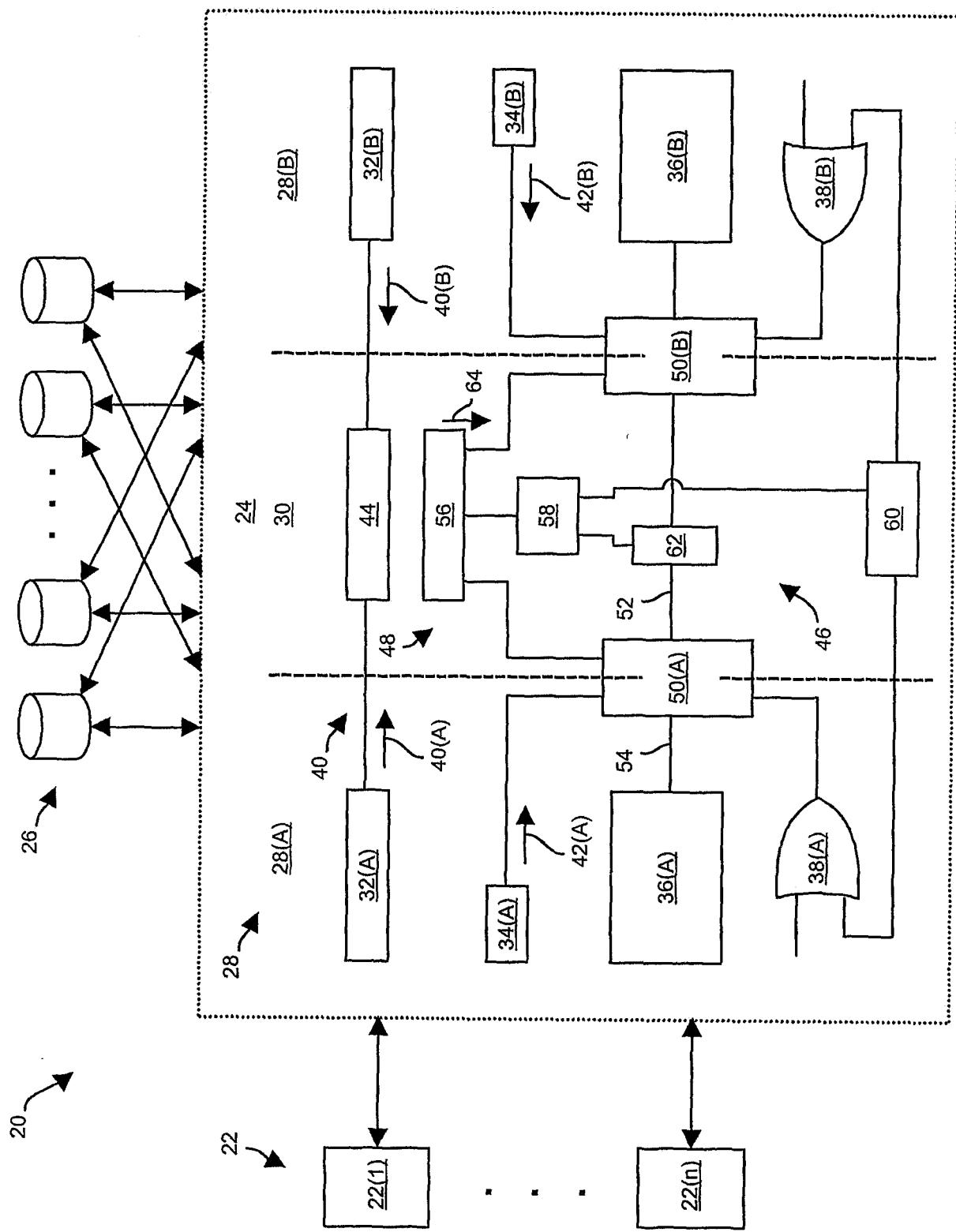


FIG. 1

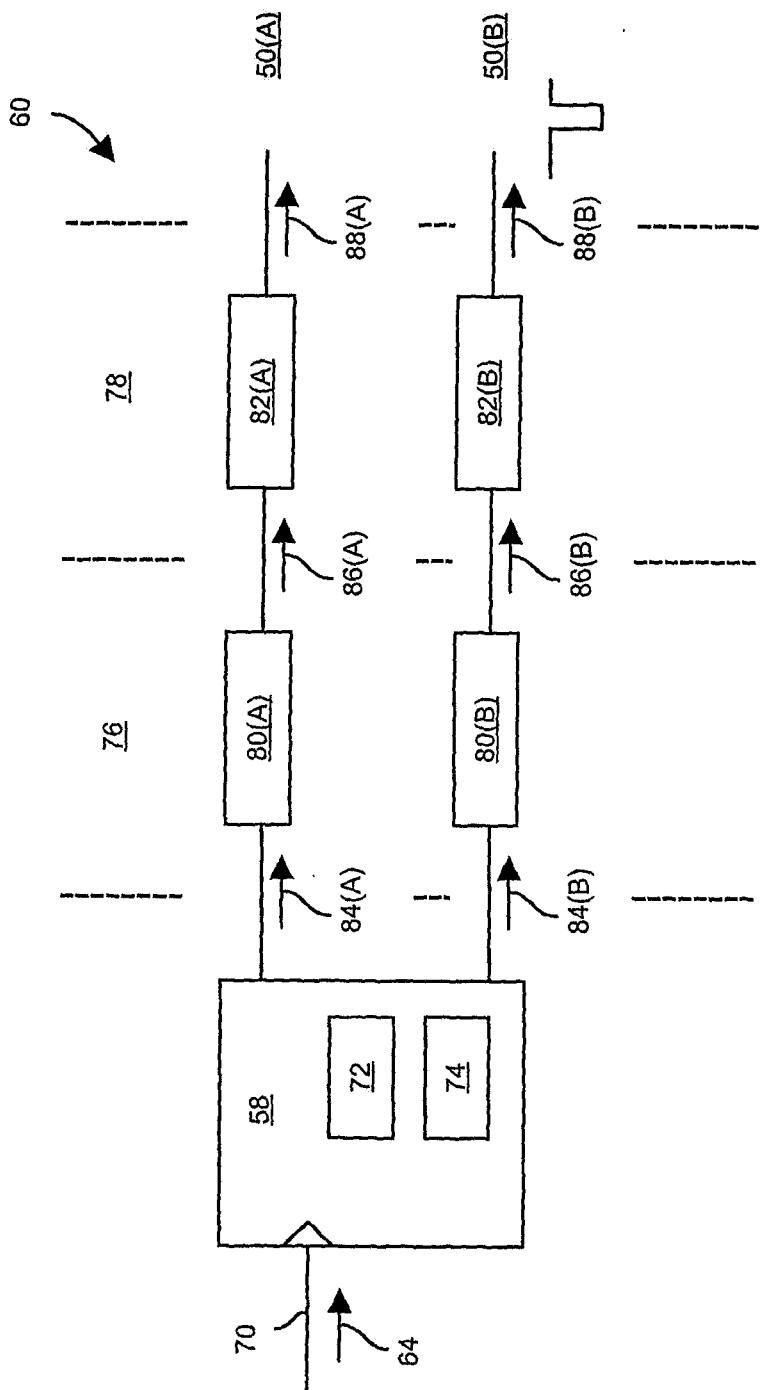


FIG. 2

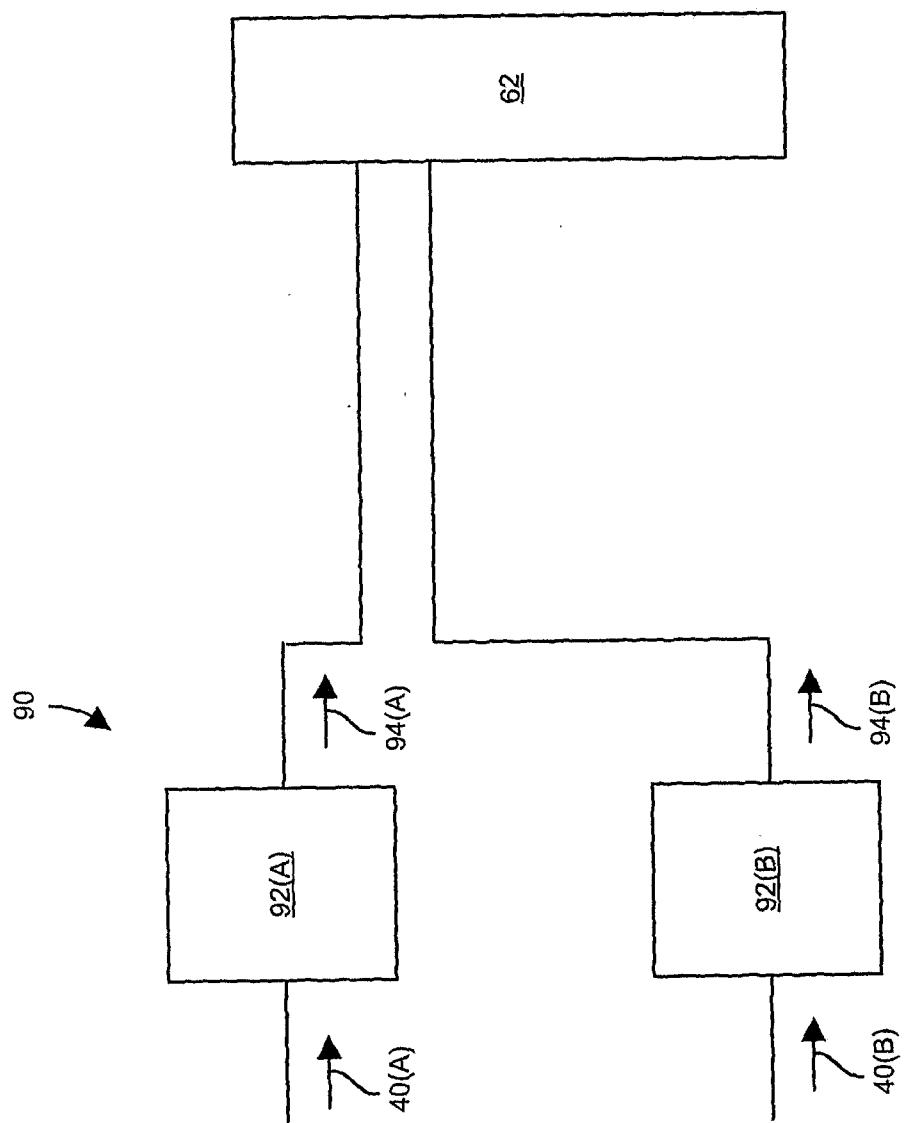


FIG. 3

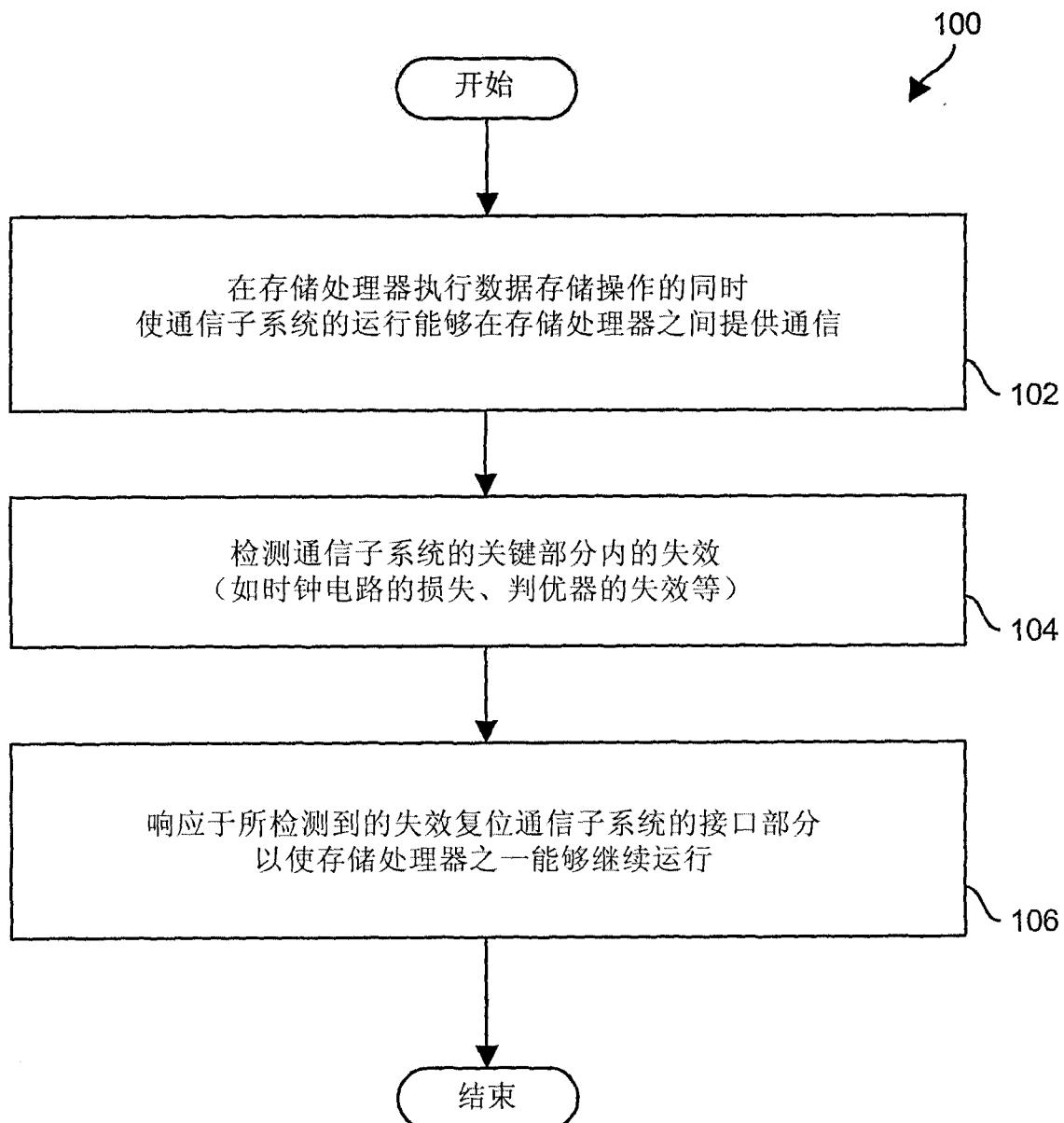


FIG. 4