[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 15/163 (2006.01)

[12] 发明专利说明书

专利号 ZL 200410057605.0

[45] 授权公告日 2007 年 1 月 10 日

[11] 授权公告号 CN 1294507C

[22] 申请日 2004.8.20
[21] 申请号 200410057605.0
[30] 优先权
[32] 2003.8.29 [33] US [31] 10/652,144
[73] 专利权人 国际商业机器公司
地址 美国纽约
[72] 发明人 肯尼思·W·博伊德
肯尼思·F·戴
菲利普·M·伯特马斯
约翰·J·沃尔夫冈
[56] 参考文献
CN1245933A 2000.3.1 G06F15/163
CN1115839C 2003.7.23 H04L29/06
CN1332924A 2002.1.23 H04L29/02
JP2003-6087A 2003.1.10 G06F13/00
WO03/024007A1 2003.3.20 H04J1/16
US6317775B1 2001.11.13 G06F15/16
审查员 尹 杰
[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 康建峰

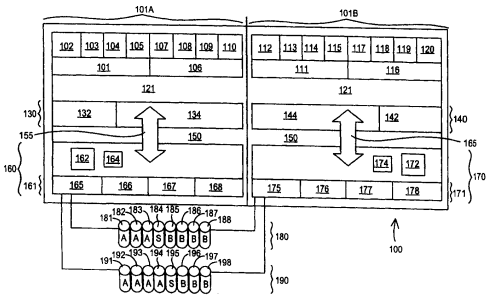
权利要求书 4 页 说明书 16 页 附图 7 页

[54] 发明名称

从多个控制节点中选择一个首领的装置和方法

[57] 摘要

一种用来从多个互连控制节点中选择一个首领控制节点的方法。该方法由该多个控制节点中的每一个提供第一信号给其他控制节点中的每一个，然后由该多个控制节点中的每一个从其他控制节点中的每一个接收响应信号。该方法然后由该多个控制节点中的每一个计算其他控制节点中的每一个的单独响应时间，并且由每个控制节点确定其合计响应时间。然后，该方法确定是否使用这些合计响应时间来选择一个首领控制节点。如果使用这些合计响应时间来选择首领控制节点，则该方法确定最小合计响应时间，并且指定具有该最小合计响应时间的控制节点为首领控制节点。



1. 一种用来从多个互连控制节点中选择首领控制节点的方法，包括以下步骤：

由所述多个控制节点中的每一个提供第一信号给其他控制节点中的每一个；

由所述多个控制节点中的每一个从其他控制节点中的每一个接收响应信号；

由所述多个控制节点中的每一个计算其他控制节点中的每一个的单独响应时间；

确定所述多个互连控制节点中的每一个的合计响应时间；

确定是否使用所述合计响应时间选择首领控制节点；

如果使用所述合计响应时间选择首领控制节点则如下操作：

确定最小合计响应时间；以及

指定具有所述最小合计响应时间的控制节点为首领控制节点。

2. 如权利要求 1 所述的方法，还包括以下步骤：

确定是否存在两个或更多个控制节点均具有所述最小合计响应时间；

如果存在两个或更多个控制节点均具有所述最小合计响应时间，则操作，从而重复如权利要求 1 所述的步骤。

3. 如权利要求 1 所述的方法，还包括以下步骤：

由所述多个控制节点中的每一个计算那个控制节点的单独响应时间的标准偏差；

由每个控制节点提供所述标准偏差给其余互连控制节点中的每一个。

4. 如权利要求 3 所述的方法，还包括以下步骤：

如果存在两个或更多个控制节点均具有最小合计响应时间，则操作，从而确定所述两个或更多个控制节点中的哪一个具有最小标准偏差；

指定具有最小合计响应时间和最小标准偏差的控制节点为首领控制节点。

5. 如权利要求 3 所述的方法，还包括以下步骤：

如果不使用所述合计响应时间选择首领控制节点，则操作，从而提供首领控制节点选择函数；

使用所述首领控制节点选择函数确定所述多个互连控制节点中的每一个的性能得分；

指定具有最小性能得分的控制节点为首领控制节点。

6. 如权利要求 5 所述的方法，其中所述首领控制节点选择函数包括以下方程：

性能得分=a(合计响应时间)^c+b(标准偏差)^d

其中 a、b、c 和 d 是正常数。

7. 如权利要求 1 所述的方法，还包括提供多个主机计算机的步骤，其中所述多个控制节点中的每一个位于所述多个主机计算机的不同之一中，并且其中所述多个主机计算机中的每一个能够与一个或多个数据存储和检索系统通信。

8. 如权利要求 1 所述的方法，还包括提供多个数据存储和检索系统的步骤，其中所述多个控制节点中的每一个位于所述多个数据存储和检索系统的不同之一中，并且其中所述多个数据存储和检索系统中的每一个能够与一个或多个主机计算机通信。

9. 如权利要求 1 所述的方法，还包括提供多个初级备份设备的步骤，其中所述多个控制节点中的每一个位于所述多个初级备份设备的不同之一中，并且其中所述多个初级备份设备中的每一个能够与一个或多个数据存储和检索系统和一个或多个次级备份设备通信。

10、一种用来从多个互连控制节点中选择首领控制节点的系统，包括以下装置：

用于由所述多个控制节点中的每一个提供第一信号给其他控制节点中的每一个的装置；

用于由所述多个控制节点中的每一个从其他控制节点中的每一个接

收响应信号的装置;

用于由所述多个控制节点中的每一个计算其他控制节点中的每一个的单独响应时间的装置;

用于确定所述多个互连控制节点中的每一个的合计响应时间的装置;

用于确定是否使用所述合计响应时间选择首领控制节点的装置;

用于确定最小合计响应时间的装置;以及

用于指定具有所述最小合计响应时间的控制节点为首领控制节点的装置。

11、如权利要求 10 所述的系统,还包括以下装置:

用于确定是否存在两个或更多个控制节点均具有所述最小合计响应时间的装置。

12、如权利要求 10 所述的系统,还包括以下装置:

用于由所述多个控制节点中的每一个计算那个控制节点的单独响应时间的标准偏差的装置;

用于由每个控制节点提供所述标准偏差给其余互连控制节点中的每一个的装置。

13、如权利要求 12 所述的系统,还包括以下装置:

用于如果存在两个或更多个控制节点均具有最小合计响应时间,则指定具有最小合计响应时间和最小标准偏差的控制节点为首领控制节点的装置。

14、如权利要求 12 所述的系统,还包括以下装置:

用于如果不使用所述合计响应时间选择首领控制节点,则使用首领控制节点选择函数确定所述多个互连控制节点中的每一个的性能得分的装置;以及

用于指定具有最小性能得分的控制节点为首领控制节点的装置。

15、如权利要求 14 所述的系统,其中所述首领控制节点选择函数包括以下方程:

性能得分=a(合计响应时间)^c+b(标准偏差)^d

其中 a、b、c 和 d 是正常数。

16、如权利要求 10 所述的系统，还包括用于提供多个主机计算机的装置，其中所述多个控制节点中的每一个位于所述多个主机计算机的不同之一中，并且其中所述多个主机计算机中的每一个能够与一个或多个数据存储和检索系统通信。

17、如权利要求 10 所述的系统，还包括用于提供多个数据存储和检索系统的装置，其中所述多个控制节点中的每一个位于所述多个数据存储和检索系统的不同之一中，并且其中所述多个数据存储和检索系统中的每一个能够与一个或多个主机计算机通信。

18、如权利要求 10 所述的系统，还包括用于提供多个初级备份设备的装置，其中所述多个控制节点中的每一个位于所述多个初级备份设备的不同之一中，并且其中所述多个初级备份设备中的每一个能够与一个或多个数据存储和检索系统和一个或多个次级备份设备通信。

从多个控制节点中选择一个首领的装置和方法

技术领域

本发明涉及一种根据性能从多个互连控制节点中选择一个首领(captain)控制节点的装置和方法。在某些实施例中，本发明涉及从多个主机计算机中选择一个首领控制节点。在某些实施例中，本发明涉及从位于多个数据存储和检索系统内的多个控制器中选择一个首领控制节点。在某些实施例中，本发明涉及从位于多个初级备份设备(primary backup appliance)内的多个控制器中选择一个首领控制节点。

背景技术

很多数据处理系统需要大量数据存储以用于高效存取、修改和再存储数据。数据存储典型地分成若干不同级别，每一个级别显现不同的数据存取时间或数据存储成本。第一或最高层数据存储涉及电子存储器，通常是动态或静态随机存取存储器(DRAM 或 SRAM)。电子存储器采取半导体集成电路的形式，其中数百万字节的数据可以存储在每个电路上，其中对这些数据字节的存取以纳秒测量。由于存取是完全电子式的，因此电子存储器提供最快的数据存取。

在某些数据处理应用中，多个有时称作“主机计算机”的互连计算机系统提供信息给多个数据存储和检索系统。从这些主机计算机中选择一个首领控制节点来协调这些主机计算机的操作将是理想的。

第二级数据存储通常涉及直接存取存储设备(DASD)。DASD 存储例如包括磁盘和/或光盘。数据比特作为盘表面上微米大小的磁性或光学改变的斑点来存储，从而表示组成数据比特二进制值的“一”和“零”。磁性 DASD 包括覆盖有残余磁性材料的一个或多个盘。这些盘旋转性地安装在受保护环境内。每个盘分成很多同心轨道或者紧密圆圈。数据沿着每个轨道逐比特地连续存储。

在某些数据处理应用中，多个主机计算机提供信息给多个互连数据

存储和检索系统。根据性能从多个数据存储和检索系统中选择一个首领控制节点来协调这些系统的操作将是理想的。

具有备份数据副本对于数据丢失将是灾难性的很多商业机构而言是强制性的。另外，还需要保护以在整个系统或者甚至是场所被诸如地震、火灾、爆炸、飓风等的灾难破坏的情况下恢复数据。

灾难恢复需要次级数据副本存储在远离于初级数据的位置上。次级场所不仅必须足够远离于初级场所，而且必须能够实时备份初级数据。当初级数据被更新时，次级场所需要实时备份初级数据，其中只有某一极小的延迟。次级场所所需的困难任务在于次级数据必须是“次序一致”的，也就是，次级数据以需要大量系统考虑的与初级数据相同的顺序次序(顺序一致性)来拷贝。顺序一致性由于在数据处理系统中存在均控制多个 DASD 的多个存储控制器而复杂化。在没有顺序一致性的情况下，将产生与初级数据不一致的次级数据，从而破坏灾难恢复。

在某些数据处理应用中，多个互连数据存储和检索系统提供数据给多个互连初级备份设备。初级备份设备形成有时所谓的一致事务集，并且周期性地将这些一致事务集提供给远程场所以进行备份存储。根据性能从多个初级备份设备中选择一个首领控制节点以协调这些备份设备的操作将是理想的。

发明内容

本发明提供一种用来从多个互连控制节点中选择首领控制节点的方法，包括以下步骤：由所述多个控制节点中的每一个提供第一信号给其他控制节点中的每一个；由所述多个控制节点中的每一个从其他控制节点中的每一个接收响应信号；由所述多个控制节点中的每一个计算其他控制节点中的每一个的单独响应时间；确定所述多个互连控制节点中的每一个的合计响应时间；确定是否使用所述合计响应时间选择首领控制节点；如果使用所述合计响应时间选择首领控制节点则如下操作：确定最小合计响应时间；以及指定具有所述最小合计响应时间的控制节点为首领控制节点。

本发明提供一种用来从多个互连控制节点中选择首领控制节点的系

统，包括以下装置：用于由所述多个控制节点中的每一个提供第一信号给其他控制节点中的每一个的装置；用于由所述多个控制节点中的每一个从其他控制节点中的每一个接收响应信号的装置；用于由所述多个控制节点中的每一个计算其他控制节点中的每一个的单独响应时间的装置；用于确定所述多个互连控制节点中的每一个的合计响应时间的装置；用于确定是否使用所述合计响应时间选择首领控制节点的装置；用于确定最小合计响应时间的装置；以及用于指定具有所述最小合计响应时间的控制节点为首领控制节点的装置。

本申请人的发明包括一种用来从多个互连控制节点中选择一个首领控制节点的装置和方法。该方法由该多个控制节点中的每一个提供第一信号给其他控制节点中的每一个，然后由该多个控制节点中的每一个从其他控制节点中的每一个接收响应信号。该方法然后由该多个控制节点中的每一个计算其他控制节点中的每一个的单独响应时间，并且由每个控制节点确定其合计响应时间。

然后，该方法确定是否使用这些合计响应时间来选择一个首领控制节点。如果使用这些合计响应时间来选择首领控制节点，则该方法确定最小合计响应时间，并且指定具有该最小合计响应时间的控制节点为首领控制节点。

附图说明

通过阅读下面结合附图的详细描述，本发明将会得到更好的理解，其中，相同的附图标记用来指定相同的单元，并且其中：

图 1 是示出本申请人的数据存储和检索系统的一个实施例的各组件的方框图；

图 2 是示出本申请人的数据存储和检索系统的第二实施例的各组件的方框图；

图 3 是示出本申请人的数据存储和检索系统的第三实施例的各组件的方框图；

图 4 是示出本申请人的对等远程拷贝数据存储和检索系统的各组件的方框图；

图 5 是概述本申请人的方法中的特定初始步骤的流程图；以及

图 6 是概述本申请人的方法中的特定附加步骤的流程图。

具体实施方式

参照附图在下面描述中以多个优选实施例描述本发明，其中相同的标号表示相同或类似的单元。本发明将被描述为实施在包括多个主机计算机、多个初级数据存储和检索系统、多个次级数据存储和检索系统和互连这些初级和次级数据存储和检索系统的多个备份设备的数据处理系统中。然而，下面对用来从多个控制节点中选择一个首领的本申请人方法的描述并不旨在将本申请人的发明限定于数据处理应用，这里的本发明还能一般应用于监视和/或协调多个计算机的操作。

图 4 示出本申请人的数据处理系统的各组件。现在参照图 4，主机计算机 480、485 和 490 通过通信链路 401 相互连接和通信。主机计算机 480、485 和 490 通过通信链路 401 与初级数据存储和检索系统 410、430 和 450 进行互连和通信。在某些实施例中，通信链路 401 从包括串行互连如 RS-232 电缆或 RS-432 电缆、以太网互连、SCSI 互连、光纤通道互连、ESCON 互连、FICON 互连、局域网(LAN)、私有广域网(WAN)、公用广域网、存储区域网(SAN)、传输控制协议/网际协议(TCP/IP)、因特网及其组合的组中选择。

主机计算机 480、485 和 490 均包括计算机系统如大型机、个人计算机、工作站等，其中包括诸如 Windows、AIX、Unix、MVS、LINUX 等的操作系统(Windows 是微软公司的注册商标；AIX 是 IBM 公司的注册商标且 MVS 是 IBM 公司的商标；而 UNIX 是通过公开组独占性地许可的在美国和其他国家的注册商标。)

计算机 480 包括一个处理器即控制节点，如控制节点 481。计算机 485 包括一个处理器即控制节点，如控制节点 486。计算机 490 包括一个处理器即控制节点，如控制节点 491。在图 4 的所示实施例中，控制节点 481、486 和 491 利用通信链路 401 相互通信。

在某些实施例中，主机计算机 480、485 和 490 分别包括存储管理程序 482、487 和 492。存储管理程序 482、487 和 492 可以包括管理向

数据存储和检索系统传输数据的本技术领域内公知的存储管理型程序的功能性，例如在 IBM MVS 操作系统中实现的 IBM DFSMS。

初级数据存储和检索系统 410 将信息从初级信息存储介质 412 提供到次级数据存储和检索系统 425 以通过初级备份设备 415 和次级备份设备 420 拷贝到次级信息存储介质 427。数据存储和检索系统 410 还包括一个处理器即控制节点 411。数据存储和检索系统 425 还包括一个处理器即控制节点 426。

在某些实施例中，信息存储介质 412 包括 DASD。在某些实施例中，信息存储介质 412 包括一个或多个 RAID 阵列。在某些实施例中，信息存储介质 412 包括多个便携式信息存储介质，例如包括多个单独位于便携式容器例如磁带盒中的磁带。

在某些实施例中，信息存储介质 427 包括 DASD。在某些实施例中，信息存储介质 427 包括一个或多个 RAID 阵列。在某些实施例中，信息存储介质 427 包括多个便携式信息存储介质，例如包括多个单独位于便携式容器例如磁带盒中的磁带。

在某些实施例中，初级备份设备 415 与初级数据存储和检索系统 410 集成在一起。在图 4 的所示实施例中，初级备份设备 415 居于初级数据存储和检索系统 410 的外部，并且通过通信链路 403 与初级数据存储和检索系统 410 通信。在某些实施例中，通信链路 403 从包括串行互连如 RS-232 电缆或 RS-432 电缆、以太网互连、SCSI 互连、光纤通道互连、ESCON 互连、FICON 互连、局域网(LAN)、私有广域网(WAN)、公用广域网、存储区域网(SAN)、传输控制协议/网际协议(TCP/IP)、因特网及其组合的组中选择。

在某些实施例中，次级备份设备 420 与次级数据存储和检索系统 425 集成在一起。在图 4 的所示实施例中，次级备份设备 420 居于次级数据存储和检索系统 425 的外部，并且通过通信链路 406 与次级数据存储和检索系统 425 通信。在某些实施例中，通信链路 406 从包括串行互连如 RS-232 电缆或 RS-432 电缆、以太网互连、SCSI 互连、光纤通道互连、ESCON 互连、FICON 互连、局域网(LAN)、私有广域网(WAN)、公用广域网、存储区域网(SAN)、传输控制协议/网际协议

(TCP/IP)、因特网及其组合的组中选择。

初级数据存储和检索系统 430 将信息从初级信息存储介质 432 提供到次级数据存储和检索系统 445 以通过初级备份设备 435 和次级备份设备 440 拷贝到次级信息存储介质 447。信息存储和检索系统 430 还包括控制节点 431。信息存储和检索系统 445 还包括控制节点 446。

在某些实施例中，信息存储介质 432 包括 DASD。在某些实施例中，信息存储介质 432 包括一个或多个 RAID 阵列。在某些实施例中，信息存储介质 432 包括多个便携式信息存储介质，例如包括多个单独位于便携式容器例如磁带盒中的磁带。

在某些实施例中，信息存储介质 447 包括 DASD。在某些实施例中，信息存储介质 447 包括一个或多个 RAID 阵列。在某些实施例中，信息存储介质 447 包括多个便携式信息存储介质，例如包括多个单独位于便携式容器例如磁带盒中的磁带。

在某些实施例中，初级备份设备 435 与初级数据存储和检索系统 430 集成在一起。在图 4 的所示实施例中，初级备份设备 435 居于初级数据存储和检索系统 430 的外部，并且通过通信链路 404 与初级数据存储和检索系统 430 通信。在某些实施例中，通信链路 404 从包括串行互连如 RS-232 电缆或 RS-432 电缆、以太网互连、SCSI 互连、光纤通道互连、ESCON 互连、FICON 互连、局域网(LAN)、私有广域网(WAN)、公用广域网、存储区域网(SAN)、传输控制协议/网际协议(TCP/IP)、因特网及其组合的组中选择。

在某些实施例中，次级备份设备 440 与次级数据存储和检索系统 445 集成在一起。在图 4 的所示实施例中，次级备份设备 440 居于次级数据存储和检索系统 445 的外部，并且通过通信链路 407 与次级数据存储和检索系统 445 通信。在某些实施例中，通信链路 407 从包括串行互连如 RS-232 电缆或 RS-432 电缆、以太网互连、SCSI 互连、光纤通道互连、ESCON 互连、FICON 互连、局域网(LAN)、私有广域网(WAN)、公用广域网、存储区域网(SAN)、传输控制协议/网际协议(TCP/IP)、因特网及其组合的组中选择。

初级数据存储和检索系统 450 将信息从初级信息存储介质 452 提供

到次级数据存储和检索系统 465 以通过初级备份设备 455 和次级备份设备 460 拷贝到次级信息存储介质 467。信息存储和检索系统 450 还包括控制节点 451。信息存储和检索系统 465 还包括控制节点 466。

在某些实施例中，信息存储介质 452 包括 DASD。在某些实施例中，信息存储介质 452 包括一个或多个 RAID 阵列。在某些实施例中，信息存储介质 452 包括多个便携式信息存储介质，例如包括多个单独位于便携式容器例如磁带盒中的磁带。

在某些实施例中，信息存储介质 467 包括 DASD。在某些实施例中，信息存储介质 467 包括一个或多个 RAID 阵列。在某些实施例中，信息存储介质 467 包括多个便携式信息存储介质，例如包括多个单独位于便携式容器例如磁带盒中的磁带。

在某些实施例中，初级备份设备 455 与初级数据存储和检索系统 450 集成在一起。在图 4 的所示实施例中，初级备份设备 455 居于初级数据存储和检索系统 450 的外部，并且通过通信链路 405 与初级数据存储和检索系统 450 通信。在某些实施例中，通信链路 405 从包括串行互连如 RS-232 电缆或 RS-452 电缆、以太网互连、SCSI 互连、光纤通道互连、ESCON 互连、FICON 互连、局域网(LAN)、私有广域网(WAN)、公用广域网、存储区域网(SAN)、传输控制协议/网际协议(TCP/IP)、因特网及其组合的组中选择。

在某些实施例中，次级备份设备 460 与次级数据存储和检索系统 465 集成在一起。在图 4 的所示实施例中，次级备份设备 460 居于次级数据存储和检索系统 465 的外部，并且通过通信链路 408 与次级数据存储和检索系统 465 通信。在某些实施例中，通信链路 408 从包括串行互连如 RS-232 电缆或 RS-452 电缆、以太网互连、SCSI 互连、光纤通道互连、ESCON 互连、FICON 互连、局域网(LAN)、私有广域网(WAN)、公用广域网、存储区域网(SAN)、传输控制协议/网际协议(TCP/IP)、因特网及其组合的组中选择。

初级备份设备 415、435 和 455 分别从初级数据存储和检索系统 410、430 和 450 接收信息。周期性地，初级备份设备 415、435 和 455 形成一致事务集。采用“一致事务集”，本申请人表示这样一个事务集，

即当在次级数据存储和检索系统控制器上应用该事务集中的所有事务时，次级存储看上去将相同于创建该事务集的时间点上的初级存储。

在某些实施例中，数据存储和检索系统 410、425、430、445、450 和/或 465 中的一个或多个包括数据存储和检索系统 100(图 1)。现在参照图 1，本申请人的信息存储和检索系统 100 包括第一群集 101A 和第二群集 101B。每个群集包括处理器部分 130/140 和输入/输出部分 160/170。每个群集的内部 PCI 总线分别通过远程 I/O 桥 155/165 连接在处理器部分 130/140 与 I/O 部分 160/170 之间。

信息存储和检索系统 100 还包括位于四个主机舱(host bay)101、106、111 和 116 内的多个主机适配器 102-105、107-110、112-115 和 117-120。每个主机适配器可以包括一个光纤通道端口、一个 FICON 端口、两个 ESCON 端口或者两个 SCSI 端口。每个主机适配器通过一个或多个公共平台互连总线 121 和 150 连接到两个群集，使得每个群集可以处理来自任何主机适配器的 I/O。

处理器部分 130 包括处理器 132 和高速缓冲存储器 134。在某些实施例中，处理器 132 包括基于 64 比特 RISC 的对称多处理器。在某些实施例中，处理器 132 包括内置故障和错误纠正功能。高速缓冲存储器 134 用来存储读取和写入数据以改善所附主机系统的性能。在某些实施例中，高速缓冲存储器 134 包括大约 4 吉字节。在某些实施例中，高速缓冲存储器 134 包括大约 8 吉字节。在某些实施例中，高速缓冲存储器 134 包括大约 12 吉字节。在某些实施例中，高速缓冲存储器 134 包括大约 16 吉字节。在某些实施例中，高速缓冲存储器 134 包括大约 32 吉字节。

处理器部分 140 包括处理器 142 和高速缓冲存储器 144。在某些实施例中，处理器 142 包括基于 64 比特 RISC 的对称多处理器。在某些实施例中，处理器 142 包括内置故障和错误纠正功能。高速缓冲存储器 144 用来存储读取和写入数据以改善所附主机系统的性能。在某些实施例中，高速缓冲存储器 144 包括大约 4 吉字节。在某些实施例中，高速缓冲存储器 144 包括大约 8 吉字节。在某些实施例中，高速缓冲存储器 144 包括大约 12 吉字节。在某些实施例中，高速缓冲存储器 144 包括大约 16 吉字节。在某些实施例中，高速缓冲存储器 144 包括大约 32 吉字

节。

I/O 部分 160 包括非易失性存储装置("NVS")162 和 NVS 电池 164。NVS 162 用来存储写入数据的第二副本以在发生群集故障的电源故障和该数据的高速缓冲存储器副本丢失的情况下确保数据完整性。NVS 162 存储提供给群集 101B 的写入数据。在某些实施例中，NVS 162 包括大约 1 吉字节的存储装置。在某些实施例中，NVS 162 包括四个独立存储卡。在某些实施例中，每对 NVS 卡具有由电池供电的充电系统，即使整个系统掉电，其也在高达 72 小时内保护数据。

I/O 部分 170 包括 NVS 172 和 NVS 电池 174。NVS 172 存储提供给群集 101A 的写入数据。在某些实施例中，NVS 172 包括大约 1 吉字节的存储装置。在某些实施例中，NVS 172 包括四个独立存储卡。在某些实施例中，每对 NVS 卡具有由电池供电的充电系统，即使整个系统断电，其也在高达 72 小时内保护数据。

在群集 101B 发生故障的情况下，故障群集的写入数据将驻留在位于正常群集 101A 内的 NVS 162 中。然后，该写入数据以高优先级降级(destage)到 RAID 等级(rank)。同时，正常群集 101A 将开始对于其自己的写入数据使用 NVS 162，从而确保仍然保持写入数据的两个副本。

I/O 部分 160 还包括多个设备适配器如设备适配器 165、166、167 和 168 和组织成两个 RAID 等级即 RAID 等级"A"和 RAID 等级"B"的十六个盘驱动器。在某些实施例中，RAID 等级"A"和"B"利用 RAID 5 协议。在某些实施例中，RAID 等级"A"和"B"利用 RAID 10 协议。

在某些实施例中，数据存储和检索系统 410、425、430、445、450 和/或 465 中的一个或多个包括数据存储和检索系统 200(图 2)。图 2 示出系统 200 的一个实施例。

系统 200 被安排用于响应来自一个或多个主机系统如主机计算机 490(图 4)的命令而存取便携式数据存储介质。系统 200 包括前壁 270 和后壁 290 上的多个存储架 260，用于存储容纳数据存储介质的便携式数据存储盒。系统 200 还包括：至少一个数据存储驱动器 250，用于对数据存储介质进行数据读取和/或写入；以及至少一个存取器(accessor)210，用于在多个存储架 260 与数据存储驱动器 250 之间运输数据存储介质。系统 200 可以可选地包括操作员面板 230 或其他用户接

口如基于万维网(web)的接口，其允许用户与存储库进行交互。系统 200 可以可选地包括上方导入/导出台 240 和/或下方导入/导出台 245，其允许将数据存储介质插入到存储库中并且/或者从存储库中移走数据存储介质而不打断存储库操作。

存取器 210 包括升降伺服部件 212，其能够沿着 Z 轴进行双向移动。存取器 210 还包括至少一个机械抓组件(gripper assembly)216，其用于抓握(gripping)一个或多个数据存储介质。在图 2 的所示实施例中，存取器 210 还包括条形码扫描器 214 或其他阅读系统如智能卡阅读器或类似系统以"阅读"有关数据存储介质的标识信息。在图 2 的所示实施例中，存取器 210 还包括位于升降伺服部件 212 上的第二机械抓机构 218。

在某些实施例中，系统 200 包括一个或多个存储框架(storage frame)，其中每一个都具有可由存取器 210 存取的存储架 260。存取器 210 在导轨(rail)205 上沿着 X 轴双向移动。在包括多个框架的存储库 100 的实施例中，这些单独框架的每一个中的导轨 205 被对齐成使得存取器 210 可以沿着邻接导轨系统从存储库的一端行驶到另一端。

在某些实施例中，数据存储和检索系统 410、425、430、445、450 和/或 465 中的一个或多个包括数据存储和检索系统 300(图 3)。现在参照图 3，虚拟磁带服务器 300("VTS")300 通过后台程序(daemon)370、372 和 374 与一个或多个主机以及一个或多个虚拟磁带服务器通信。在图 3 的所示实施例中，后台程序 370 通过通信链路 380 与第一主机通信。在图 3 的所示实施例中，后台程序 372 通过通信链路 382 与第二主机通信。后台程序 374 通过通信链路 384 与例如初级备份设备如设备 415 通信。

VTS 300 还与直接存取存储设备(DASD)310、多个数据存储设备 330 和 340 通信。在某些实施例中，数据存储设备 330 和 340 位于一个或多个数据存储和检索系统内。在某些实施例中，DASD 310 与主机 110 集成在一起(图 1)。在某些实施例中，DASD 310 与 VTS 300 集成在一起。在某些实施例中，DASD 310 与数据存储和检索系统集成在一起。在某些实施例中，DASD 310 外部于主机 110、VTS 300 以及与 VTS 300 通信的一个或多个数据存储和检索系统。

VTS 300 还包括存储管理器 320 如 IBM Adstar®分布式存储管理器。存储管理器 320 控制从 DASD 310 到安装在数据存储设备 330 和 340 中的信息存储介质的数据移动。在某些实施例中，存储管理器 320 包括 ADSM 服务器 322 和 ADSM 分级式存储管理器客户端 324。可替换地，服务器 322 和客户端 324 均可包括 ADSM 系统。来自 DASD 310 的信息通过 ADSM 服务器 322 和 SCSI 适配器 385 提供到数据存储设备 330 和 340。

VTS 300 还包括存储管理器 320 如 IBM Adstar®分布式存储管理器。存储管理器 320 控制从 DASD 310 到安装在数据存储设备 330 和 340 中的信息存储介质的数据移动。在某些实施例中，存储管理器 320 包括 ADSM 服务器 322 和 ADSM 分级式存储管理器客户端 324。可替换地，服务器 322 和客户端 324 均可包括 ADSM 系统。来自 DASD 310 的信息通过 ADSM 服务器 322 和 SCSI 适配器 385 提供到数据存储设备 330 和 340。

VTS 300 还包括自主控制器 350。自主控制器 350 通过分级式存储管理器(HSM)客户端 324 控制 DASD 310 的操作，以及 DASD 310 与数据存储设备 330 和 340 之间的数据传输。

回到图 4，每个主机计算机提供信息给一个或多个初级数据存储和检索系统。为了最大化利用通信链路 401 的带宽，主机计算机 480、485 和 490 必须交互以分配该带宽。在本申请人方法的某些实施例中，控制节点 481、486 和 491 交互以根据性能选择首领控制节点。该首领主机控制节点协调特定功能，例如由每一个主机计算机 480、485 和 490 形成一致事务集。序列号为 10/339,957、名称为“Method、System and Article of Manufacture for Creating a Consistent Copy”且转让给其共同受让人的未决专利申请描述了一种形成一致事务集的方法，并且在此将其全文引作参考。

回到图 4，每个初级数据存储和检索系统 410、430 和 450 以不同数据传输速率从不同主机计算机接收不同数量的信息。每个初级数据存储和检索系统 410、430 和 450 以不同数据传输速率将不同数量的信息提供到初级备份设备 415、435 和 455 中的一个或多个。在本申请人方法的某些实施例中，控制节点 411、431 和 451 交互以根据性能选择首领

控制节点。该首领主机控制节点协调特定功能，例如由每一个初级数据存储和检索系统 410、430 和 450 形成一致事务集。

回到图 4，每个初级备份设备以不同于其他初级备份设备的速率从不同初级存储控制节点接收数据。在本申请人方法的某些实施例中，控制节点 417、437 和 457 交互以根据性能选择首领控制节点。该首领备份设备控制节点协调特定功能，例如由每一个初级备份设备 415、435 和 455 形成一致事务集。

初级备份设备如设备 415、435 和 455 通过公共通信链路如通信链路 409 分别提供一致事务集到其对应的次级备份设备如设备 420、440 和 460。在某些实施例中，通信链路 409 从包括串行互连如 RS-232 电缆或 RS-432 电缆、以太网互连、SCSI 互连、光纤通道互连、ESCON 互连、FICON 互连、局域网(LAN)、私有广域网(WAN)、公用广域网、存储区域网(SAN)、传输控制协议/网际协议(TCP/IP)、因特网及其组合的组中选择。

一般而言，本申请人的方法包括用于根据实际性能标准从多个互连控制节点中选择一个首领控制节点的方法。图 5 概述了本申请人的方法的特定步骤。现在参照图 5，在步骤 505，该方法提供多个即总共(N)个互连控制节点，其中(N)大于或等于 2。在某些实施例中，(N)个互连控制节点中的每一个位于(N)个主机计算机的不同之一中。在某些实施例中，(N)个互连控制节点中的每一个位于(N)个数据存储和检索系统的不同之一中。在某些实施例中，(N)个互连控制节点中的每一个位于(N)个备份设备的不同之一上。

在步骤 510，每个控制节点在第一时间将第一信号提供给其他(N-1)个互连控制节点中的每一个。本申请人的方法从步骤 510 转至步骤 515，其中当接收到第一控制信号时，(N)个控制节点中的每一个都提供响应信号。因此，步骤 515 包括由(N)个控制节点中的每一个提供(N-1)个响应信号。

本申请人的方法从步骤 515 转至步骤 520，其中在(N-1)个第二时间，每个控制节点从其他(N-1)个控制节点接收响应信号。本申请人的方法从步骤 520 转至步骤 525，其中每个控制节点计算(N-1)个对其心跳(heart beat)信号的单独响应时间，即其他(N-1)个控制节点中每一个的响应时间。本领域的技术人员应当理解，控制节点通过从步骤 510 的第一

时间减去步骤 520 的第(i)个第二时间来计算第(i)其他控制节点的响应时间。

本申请人的方法从步骤 525 转至步骤 530，其中每个控制节点确定其合计响应时间。第一控制节点的合计响应时间包括其他控制节点响应第一控制节点的心跳信号的(N-1)个响应时间之和。在某些实施例中，步骤 530 还包括由每个控制节点向其他控制节点中的每一个报告其合计响应时间。该报告可以包括本领域的技术人员公知的任何信号通知方法。例如，每个控制节点可以发送包括其合计响应时间的消息到其他(N-1)个控制节点中的每一个。在其他实施例中，每个控制节点轮询其他(N-1)个控制节点以从这些其他控制节点获得合计响应时间。在某些实施例中，本申请人的方法从步骤 530 转至步骤 540。

在某些实施例中，本申请人的方法从步骤 530 转至步骤 535，其中每个控制节点计算(N-1)个包括其合计响应时间的单独响应时间的标准偏差。本领域的技术人员应当理解，标准偏差表示所有数据点的分布聚集在平均值周围的紧密程度。当数据紧密地聚集在一起即钟形曲线陡峭时， 标准偏差小。当数据分散开来且钟形曲线较平坦时， 标准偏差较大。

在某些实施例中，步骤 535 还包括由每个控制节点向其他互连控制节点中的每一个报告其标准偏差。该报告可以包括本领域的技术人员公知的任何信号通知方法。例如，每个控制节点可以发送包括步骤 530 的标准偏差的消息到其他(N-1)个控制节点中的每一个。在某些实施例中，每个控制节点在步骤 535 轮询其他(N-1)个控制节点以从这些其他控制节点获得标准偏差。

本申请人的方法从步骤 535 转至步骤 540，其中该方法确定是否使用(N)个合计响应时间作为主要决定因素来选择首领控制节点。如果本申请人的方法在步骤 540 决定使用(N)个合计响应时间作为主要决定因素来选择首领控制节点，则该方法从步骤 540 转至步骤 545，其中(N)个控制节点中的每一个独立识别最小合计响应时间。

本申请人的方法从步骤 545 转至步骤 550，其中每个控制节点确定

是否存在两个或更多个控制节点具有最小合计响应时间。如果本申请人的方法在步骤 550 确定有两个或更多个控制节点具有最小合计响应时间，则该方法从步骤 550 转至步骤 610(图 6)。如果本申请人的方法在步骤 550 确定不存在两个或更多个控制节点具有最小合计响应时间，则该方法从步骤 550 转至步骤 555，其中该方法指定具有最小合计响应时间的控制节点为首领控制节点。

本申请人的方法从步骤 555 转至步骤 570，其中该方法确定是否到达选择新首领控制节点的时间。如果本申请人的方法在步骤 570 确定尚未达到选择新首领控制节点的时间，则该方法周期性地返回到步骤 570。如果本申请人的方法在步骤 570 确定到达选择新首领控制节点的时间，则该方法从步骤 570 转至步骤 510 并且继续。

如果本申请人的方法在步骤 540 决定不使用(N)个合计响应时间作为主要决定因素来选择首领控制节点，则该方法从步骤 540 转至步骤 560，其中本申请人的方法提供首领控制节点选择函数。在某些实施例中，步骤 560 的首领控制节点选择函数设置在位于主机计算机内的固件如固件 481(图 4)、486(图 4)、491(图 4)中。在某些实施例中，步骤 560 的首领控制节点选择函数设置在位于数据存储和检索系统内的固件如固件 412(图 4)、432(图 4)、452(图 4)中。在某些实施例中，步骤 560 的首领控制节点选择函数设置在位于备份设备内的固件如固件 416(图 4)、436(图 4)、456(图 4)中。

步骤 560 还包括利用首领控制节点选择函数确定每一个互连控制节点的性能得分。在某些实施例中，每个控制节点根据步骤 530 的合计响应时间和步骤 535 的标准偏差为(N)个控制节点中的每一个计算性能得分。在某些实施例中，步骤 560 包括使用方程(1)确定每个控制节点的性能得分:

性能得分=a(合计响应时间)^c+b(标准偏差)^d (1)

其中 a、b、c 和 d 是正常数。较低性能得分包括较佳性能，即较小合计响应时间和较小标准偏差。

本申请人的方法从步骤 560 转至步骤 565，其中该方法指定具有步

骤 560 的最小性能得分的控制节点为首领控制节点。在某些实施例中，所有控制节点独立执行步骤 560 和 565 以指定具有最小性能得分的控制节点为首领控制节点。本申请人的方法从步骤 565 转至步骤 570。

如果本申请人的方法在步骤 550 确定存在两个或更多个控制节点具有最小合计响应时间，则该方法转至步骤 610(图 6)，其中该方法确定是否重复图 5 的步骤。在某些实施例中，如果存在两个或更多个控制节点具有最小合计响应时间则是否重复图 5 的步骤设置在位于主机计算机内的固件如固件 481(图 4)、486(图 4)、491(图 4)中。在某些实施例中，如果存在两个或更多个控制节点具有最小合计响应时间则是否重复图 5 的步骤设置在位于备份设备内的固件如固件 416(图 4)、436(图 4)、456(图 4)中。

如果本申请人的方法决定重复图 5 的步骤，则该方法从步骤 610 转至步骤 510 并且继续。如果本申请人的方法决定不重复图 5 的步骤，则该方法从步骤 610 转至步骤 620，其中该方法确定是否存在一个具有最小合计响应时间的控制节点具有小于任何其他具有最小合计响应时间的控制节点的标准偏差。如果本申请人的方法确定存在一个具有最小合计响应时间的控制节点具有小于任何其他具有最小合计响应时间的控制节点的标准偏差，则该方法从步骤 620 转至步骤 630，其中每个控制节点指定具有最小合计响应时间和较小标准偏差的控制节点为首领控制节点。本申请人的方法从步骤 630 转至步骤 570 并且继续。

如果本申请人的方法在步骤 620 确定不存在一个具有最小合计响应时间的控制节点具有小于任何其他具有最小合计响应时间的控制节点的标准偏差，则该方法从步骤 620 转至步骤 640，其中该方法指定具有最小合计响应时间的控制节点之一为首领控制节点。本申请人的方法从步骤 640 转至步骤 570 并且继续。

在某些实施例中，图 5 和/或 6 所述的各个步骤可以经过组合、删除或重新排序。

本申请人的发明还包括一种制造品，其包括一个计算机可用介质例如计算机可用介质 413、418、433、438、453、458、483、488 和/或

493，其中包含有用来使用图 5 和/或 6 所述的步骤选择首领控制节点的计算机可读程序代码。

本申请人的发明还包括一种可与可编程计算机处理器一起使用的计算机程序产品，例如计算机程序产品 414、419、434、439、454、459、484、489 和/或 494，其具有使用图 5 和/或 6 所述的步骤选择首领控制节点的计算机可读程序代码。

虽然详细地举例说明了本发明的优选实施例，但是本领域的技术人员应当清楚，在不脱离由所附权利要求限定的本发明的范围的情况下，可以对这些实施例进行各种修改和变动。

图 1

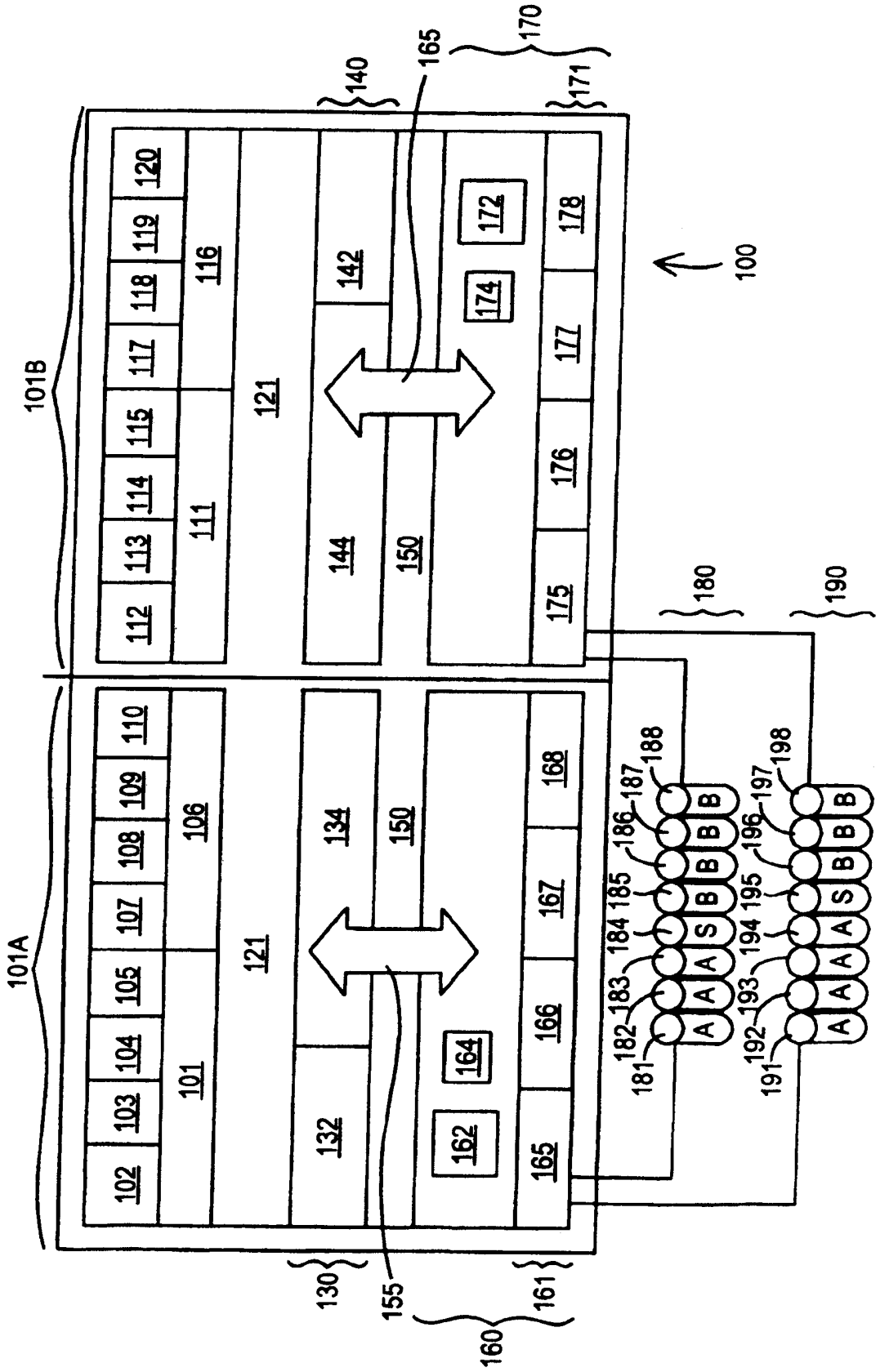

101A
101B
100
102
103
104
105
106
107
108
109
110
101
111
112
113
114
115
116
117
118
119
120
121
130
132
134
140
142
144
150
155
160
161
162
164
165
166
167
168
170
171
172
174
175
176
177
178
180
181
182
183
184
185
186
187
188
190
191
192
193
194
195
196
197
198
A
S
B

图 2

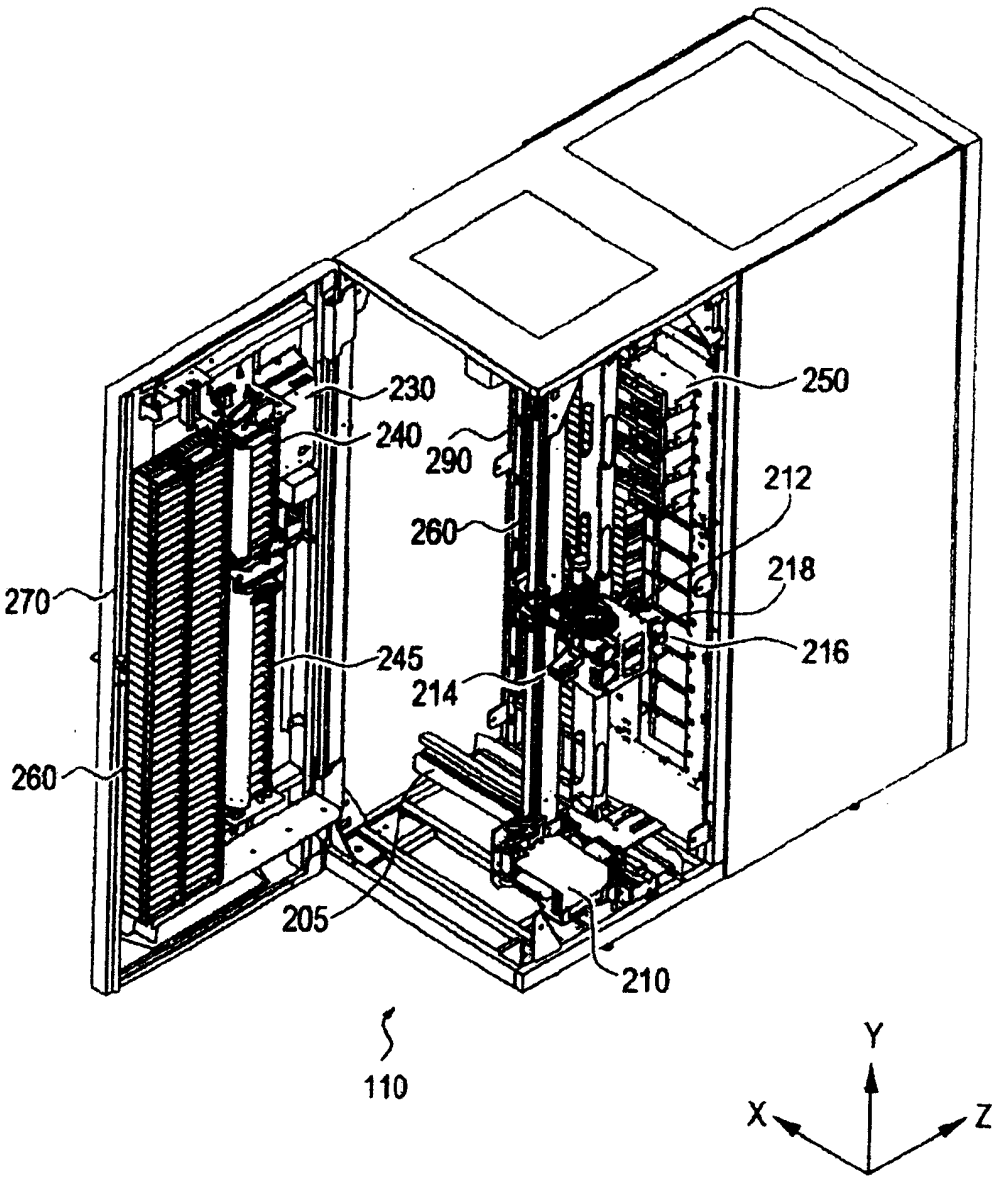


图 3

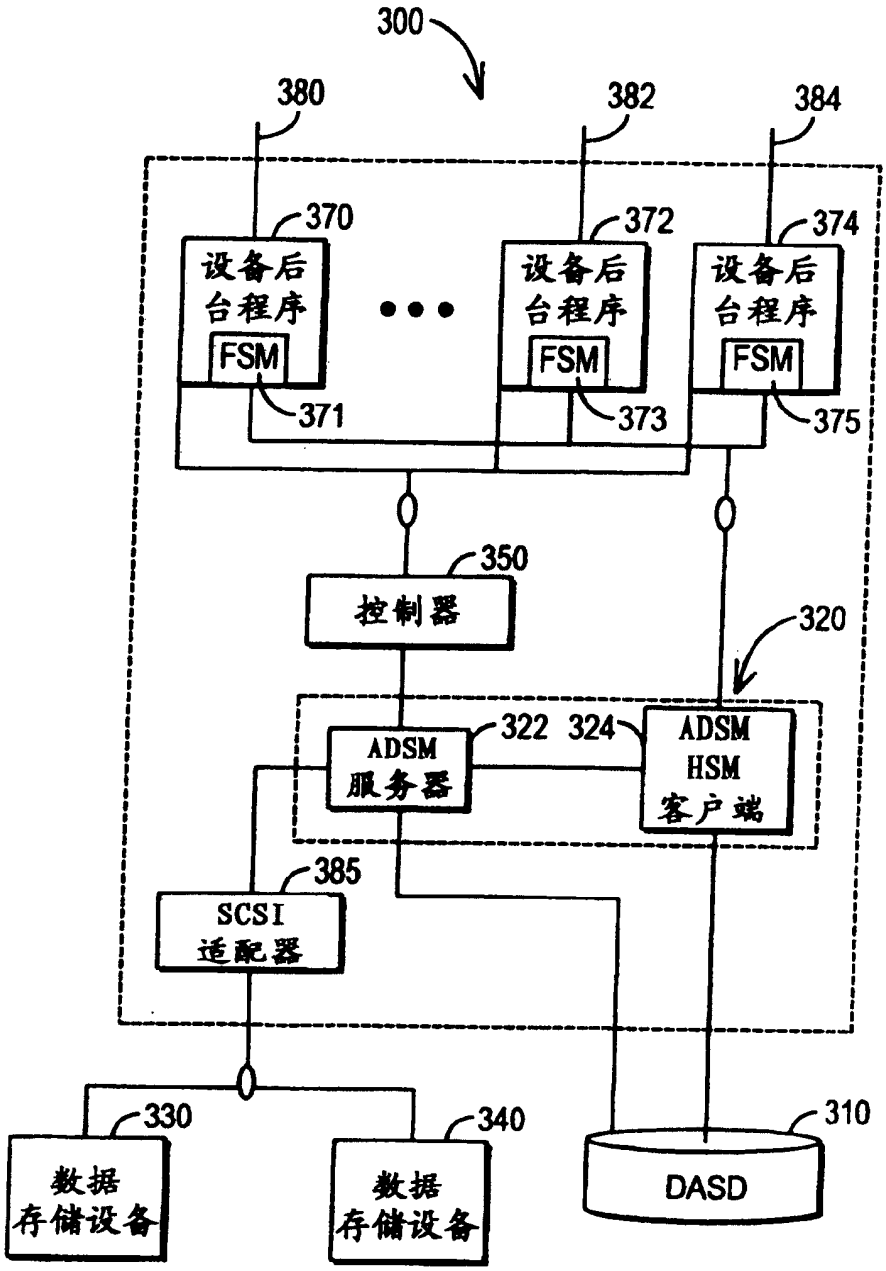
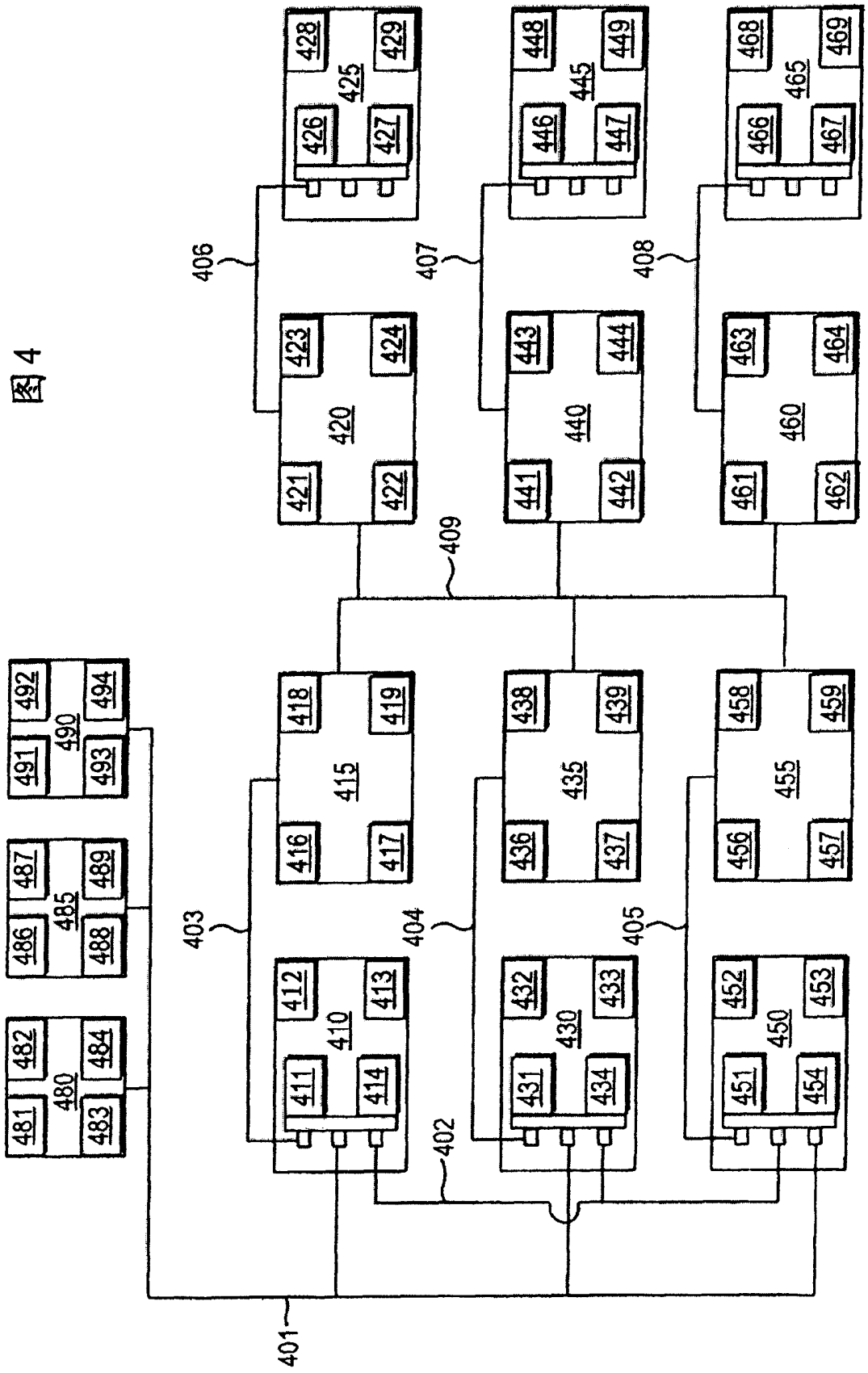

300
380
382
384
370
372
374
设备后台程序
FSM
设备后台程序
FSM
设备后台程序
FSM
371
373
375
350
控制器
320
ADSM服务器
322
324
ADSM HSM客户端
385
SCSI适配器
330
340
310
数据存储设备
数据存储设备
DASD

图 4


401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494

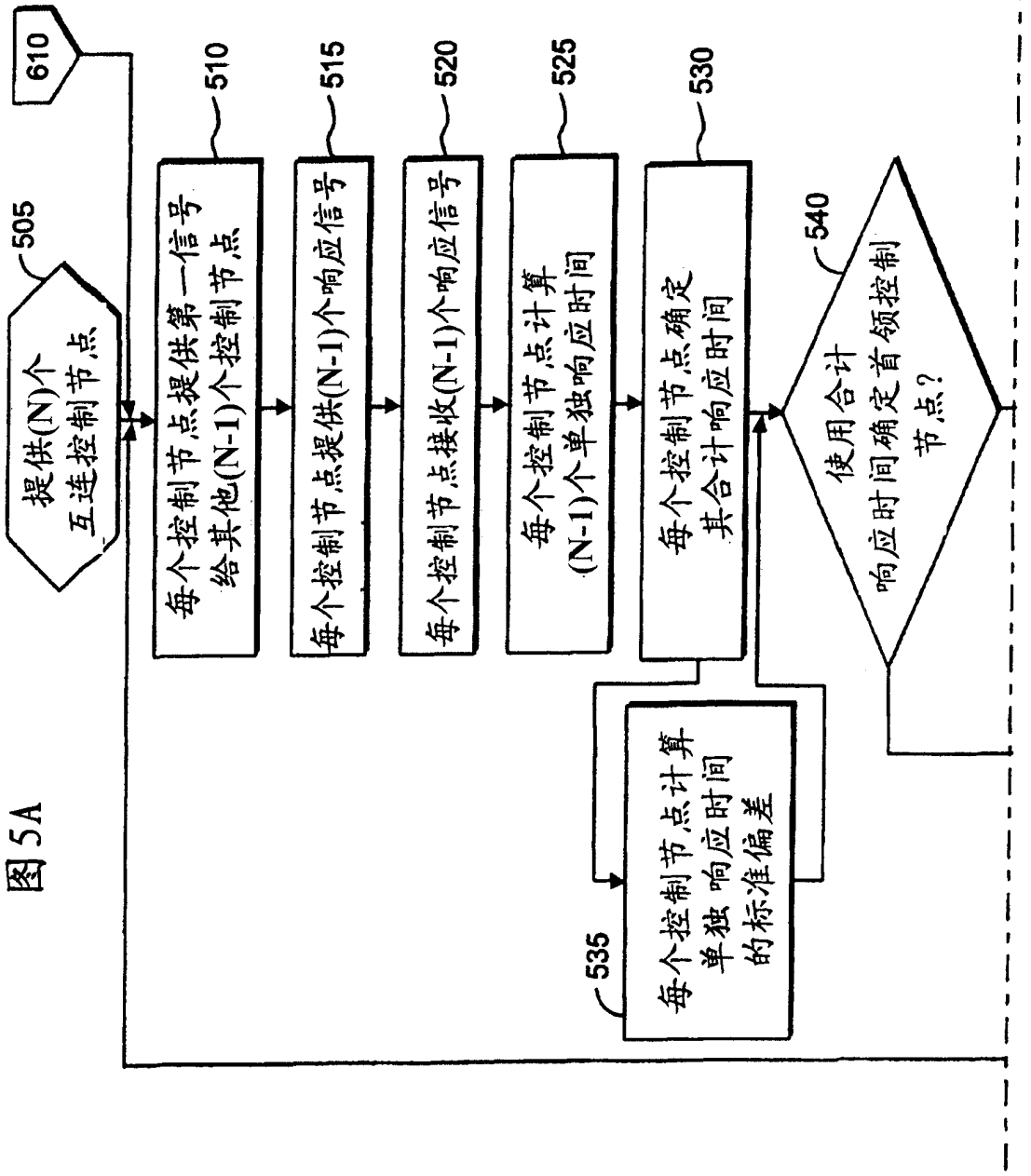

610
505
提供(N)个
互连控制节点
510
每个控制节点提供第一信号
给其他(N-1)个控制节点
515
每个控制节点提供(N-1)个响应信号
520
每个控制节点接收(N-1)个响应信号
525
每个控制节点计算
(N-1)个单独响应时间
530
每个控制节点确定
其合计响应时间
540
使用合计
响应时间确定首领控制
节点?
535
每个控制节点计算
单独响应时间
的标准偏差

图5A

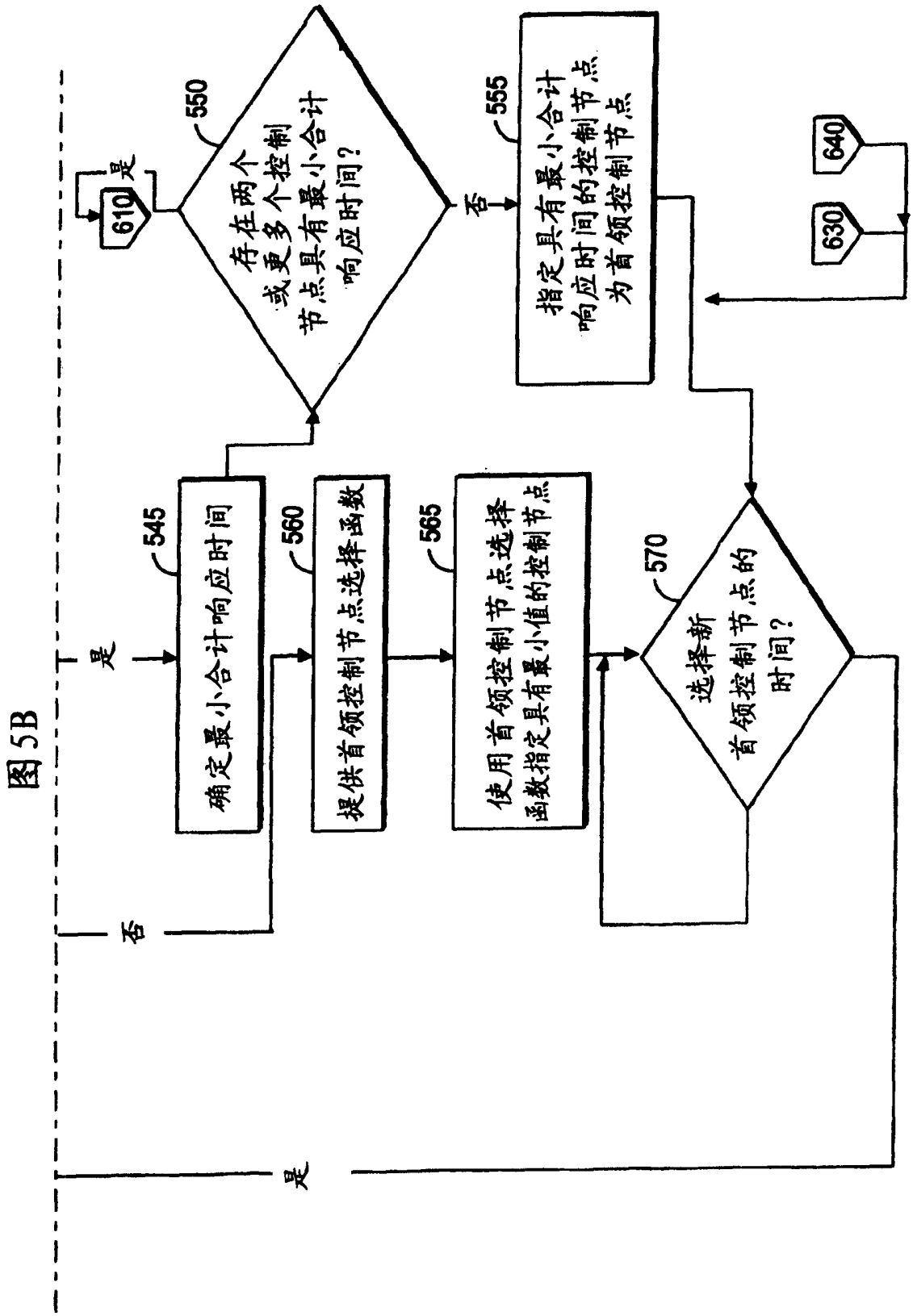
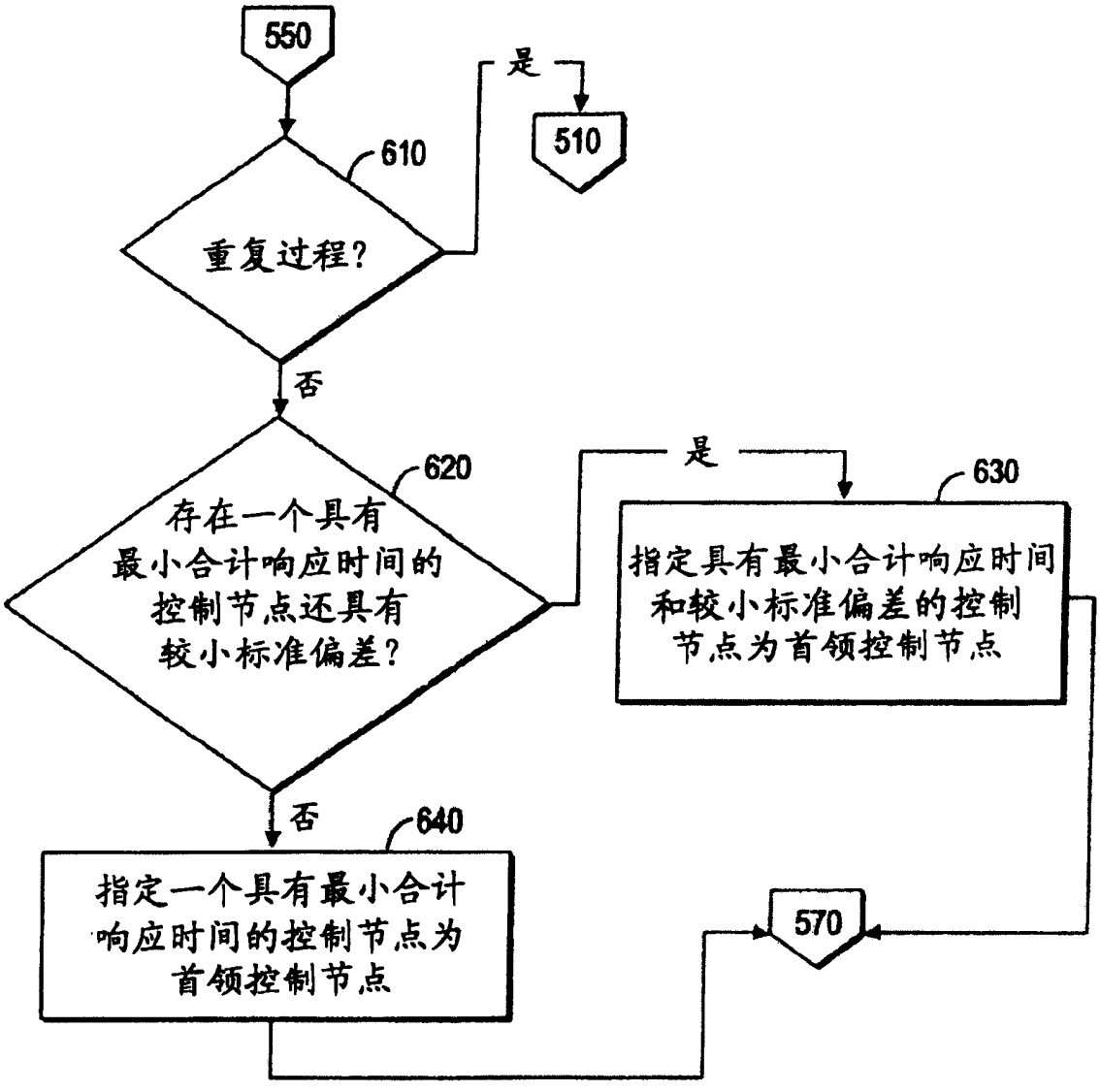

610
是
550
存在两个
或更多个控制
节点具有最小合计
响应时间?
否
555
指定具有最小合计
响应时间的控制节点
为首领控制节点
640
630
是
545
确定最小合计响应时间
否
560
提供首领控制节点选择函数
565
使用首领控制节点选择
函数指定具有最小值的控制节点
570
选择新
首领控制节点的
时间?
是

图5B

图6


550
是
510
610
重复过程?
否
620
存在一个具有
最小合计响应时间的
控制节点还具有
较小标准偏差?
是
630
指定具有最小合计响应时间
和较小标准偏差的控制
节点为首领控制节点
否
640
指定一个具有最小合计
响应时间的控制节点为
首领控制节点
570