[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04L 12/26 (2006.01)
G06F 11/36 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200610121617.4

[43] 公开日 2007 年 3 月 14 日

[11] 公开号 CN 1929412A

[22] 申请日 2006.8.23
[21] 申请号 200610121617.4
[30] 优先权
[32] 2005. 9. 9 [33] US [31] 11/223,887
[71] 申请人 国际商业机器公司
地址 美国纽约
[72] 发明人 威廉·M·萨卡尔
丹尼斯·M·塞维格尼
斯科特·M·卡尔森
多纳德·克拉特里
特雷沃·E·卡尔森
乔纳桑·D·布拉巴利
戴维·A·埃尔克
米切尔·亨利·索多尔·哈克
老罗纳德·M·斯密斯 张 立

[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 李 颖

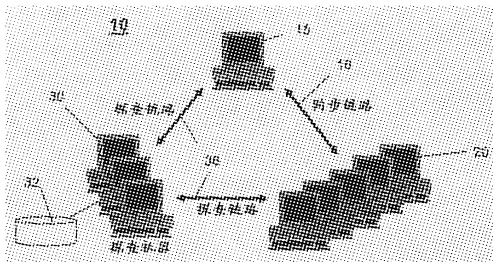
权利要求书 3 页 说明书 9 页 附图 3 页

[54] 发明名称

多节点网络的动态调试设备

[57] 摘要

动态调试多节点网络的系统、方法和计算机程序产品，所述多节点网络包含包括多个设备的基础结构，每个设备适合于在节点之间传递消息，所述消息可包括使设置在每个节点的定时时钟同步的信息。 所述设备包括使每个节点与一个探查装置互连的多个探查链路，所述探查装置监视包括在由节点传递的每条消息中的数据。 每个探查装置处理来自每条消息的数据，从而确定触发条件在节点处的存在，并且响应检测到触发条件，产生由网络中的所有节点接收的专用消息。 每个节点通过中断节点处的操作和记录可用于调试的数据，响应所述专用消息。 这样，在首次错误检测时在每个节点收集调试信息，并在执行时间动态收集调试信息，而不必手动干预。



1、一种用于对多节点网络进行动态调试的设备，所述网络包含包括多个设备的基础结构，每个设备适合于按照协议在节点之间传递消息，所述设备包括:

使所述多节点网络的每个节点与探查装置互连的多个探查链路，所述探查装置监视包括在由所述网络中的节点传递的每条消息中的数据;

在所述探查装置处理来自每条消息的数据，从而确定触发条件在节点处的存在，并且响应检测到触发条件，产生通过所述探查链路由所述网络中的所有节点接收的消息，以中断所述节点处的操作并记录可用于调试的数据的装置，从而在首次错误检测时在每个节点收集调试信息，并在执行时间动态收集调试信息，而不必手动干预。

2、按照权利要求 1 所述的设备，其中针对由所述协议定义的在范围之外的值定义触发条件。

3、按照权利要求 1 所述的设备，其中针对由所述协议定义的多个消息字段内容间的不一致值，定义触发条件。

4、按照权利要求 1 所述的设备，其中定义触发条件以检测由所述协议定义的网络变化的有效性。

5、按照权利要求 1 所述的设备，其中探查装置计算相对于它正在探查的节点设备的协议相关值，并在多个节点设备间比较这些值，其中针对与比较所述值的结果相关的情况定义所述触发条件。

6、按照权利要求 1 所述的设备，其中所述协议管理用于使网络中的系统时钟同步的定时信息的传递，所述设备被用于检测设置在节点设备的系统时钟之间的长期时钟漂移率。

7、按照权利要求 1 所述的设备，其中所述协议管理用于使网络中的系统时钟同步的定时信息的传递，所述设备被用于检测在运行相同或不同同步算法的定时网络内的系统时钟方面的差异。

8、按照权利要求 1 所述的设备，其中所述协议管理用于使网络

中的系统时钟同步的定时信息的传递，所述设备被用于确定在各种不利的环境下，不同的同步算法的行为。

9、一种用于对多节点网络进行动态调试的方法，所述网络包含包括多个设备的基础结构，每个设备适合于按照协议在节点之间传递消息，所述方法包括：

实现通过探查链路监视由所述网络中的节点传递的每条消息中包括的数据的探查装置；

在所述探查装置处理来自每条消息的数据，从而确定触发条件在节点处的存在，并且响应检测到触发条件，

产生通过所述探查链路由所述网络中的所有节点接收的消息，用于中断所述节点处的操作并记录可用于调试的数据，从而在首次错误检测时在每个节点收集调试信息，并在执行时间动态收集调试信息，而不必手动干预。

10、按照权利要求9所述的方法，其中所述处理数据包括确定由所述协议定义的在范围之外的值，所述在范围之外的值定义触发条件。

11、按照权利要求9所述的方法，其中所述处理数据包括确定由所述协议定义的多个消息字段内容间的不一致值，所述多个消息字段内容间的不一致值定义触发条件。

12、按照权利要求9所述的方法，其中所述处理数据包括检测由所述协议定义的网络变化的有效性，无效的网络变化定义触发条件。

13、按照权利要求9所述的方法，其中所述探查装置计算相对于它正在探查的节点设备的协议相关值，并在多个节点设备间比较这些值，其中针对与比较所述值的结果相关的情况定义所述触发条件。

14、按照权利要求9所述的方法，其中所述协议管理用于使网络中的系统时钟同步的定时信息的传递，所述方法包括：检测设置在节点设备的系统时钟之间的长期时钟漂移率。

15、按照权利要求9所述的方法，其中所述协议管理用于使网络中的系统时钟同步的定时信息的传递，所述方法包括：检测在运行相同或不同同步算法的定时网络内的系统时钟方面的差异。

16、按照权利要求9所述的方法，其中所述协议管理用于使网络中的系统时钟同步的定时信息的传递，所述方法包括：确定在各种不利的环境下，不同的同步算法的行为。

17、一种机器可读的程序存储装置，所述程序存储装置确实包含可由机器执行从而实现动态调试多节点网络的方法步骤的指令的程序，所述网络包含包括多个设备的基础结构，每个设备适合于按照协议在节点之间传递消息，所述方法是按照前述方法权利要求任意之一所述的方法。

多节点网络的动态调试设备

技术领域

本发明涉及多处理器/多节点网络，尤其涉及动态监视在节点间传送的网络分组，检测错误和实时收集在节点的调试数据的系统和方法。

背景技术

目前，在多处理器/多节点连网系统中，一个处理器可能发生问题，但是由于监视系统的监视工具/人员的反应缓慢，以及网络中所包含的节点的数目的缘故，支持该问题的调试（debugging）的数据可能丢失。如果为了调试该问题，需要来自在许多位置的许多节点/系统的带有同时的时戳的数据，那么该问题被进一步复杂化。

美国专利 No.5119377 描述一种用于连网计算机系统的问题管理系统，藉此，在软件开发期间，检错码被放置在软件程序内。当在节点检测到错误或故障时，执行只捕捉调试软件错误所需的数据的处理。在执行时间之前静态地定义待捕捉的数据。

美国专利 No.6769077 描述一种远程内核调试系统，其中主计算机调试器远程发出停止目标计算机的核心操作系统的执行的命令，并且通过串行总线由主计算机提取和保存目标计算机的物理存储器的快照。

对该问题的其它解决方案包括诸如被动地监视网络分组的局域网(LAN)“嗅探器”之类的产品。它们借助过滤器定义，收集特定类型的网络分组。这种解决方案的缺陷在于数据缓冲器会很快溢出，感兴趣的分组会丢失。此外，这种解决方案并不基于问题出现触发数据收集。另一缺陷在于任何分组分析需要数据的后处理。

理想的是提供一种收集和在连网计算机的节点处发生的程序故

障有关的信息的系统和方法，其中在首次错误检测时收集调试信息，并在执行时间动态地从多个系统收集调试信息。

发明内容

本发明的目的在于一种监视许多远程节点/系统，检查在所涉及的系统之间传递的数据，如果检测到错误，那么对网络中的所有系统发送保存其当前状态/调试数据的命令的方法和系统。

本发明的方法和系统包括一个或多个探查装置（probe device），所述一个或多个探查装置监视许多远程节点/系统间的分组通信，并当分组在探查装置被收集时，实时地处理所述分组。这样，和 LAN 嗅探器的情况不一样，系统不是被动的数据收集器。特别地，探查装置实时地检查数据，并且一旦检测到错误，就触发在远程节点的数据收集，保证在本地和远程节点所有必需的调试数据都被收集。可对于正被监视的每种条件，逐个情况地确定该数据。

根据本发明的一个方面，提供一种动态调试包含包括多个设备的基础结构的多节点网络的系统、方法和计算机程序产品，每个设备适合于通过节点之间的链路传递消息，包括使设置在每个节点中的定时时钟同步的信息。所述设备包括使每个节点和监视包括在由节点传递的每条消息中的数据的探查装置互连的探查链路。每个探查装置处理来自每条消息的数据，确定触发条件在节点处的存在，响应检测到触发条件，产生通过探查链路由网络中的所有节点接收的指令每个节点收集有关的调试数据和/或中断操作的消息。这样，在首次错误检测时，在每个节点收集调试信息，并在执行时间动态收集调试信息，而不必手动干预。

有利的是，本发明的系统和方法允许从多个系统同时收集调试数据。此外，本发明可用在实现其它协议，从而在计算机设备的互连网络的节点间传递信息，以便调试这样的节点的系统中。可根据被分析的协议定义触发。

附图说明

结合附图，根据下面的详细说明，对本领域的技术人员来说，本发明的目的、特征和优点将变得明显，其中：

图 1 是描述其中实现本发明的系统 10 的图；

图 2 是图解说明根据本发明，由定时网络中的探查装置实现的方法 50 的流程图；

图 3A-3B 图解说明根据本发明，检测网络中的各种触发条件之一，并在网络的节点启动中断和数据保存功能的步骤。

具体实施方式

本发明的目的在于一种在首次错误检测时，收集和异常有关的信息，并在执行时间动态收集所述信息的系统和方法。

图 1 图解说明实现收集和在连网计算机的节点处发生的程序故障有关的信息的系统和方法的连网系统 10，其中在首次错误检测时收集调试信息，并在执行时间动态收集所述调试信息。在一个实施例中，系统 10 包括一个或多个服务器设备 15 或客户机设备 20，这里，每个设备 15、20 被交替称为中央电子复合体(CEC)，中央电子复合体一般意味封装成单一实体的处理器复合体。例如，一个 CEC 15、20 可包含 IBM 系统 z9，或者 IBM eServer® zSeries®(例如 zSeries 990(z990)或 zSeries 890(z890))系统。在图 1 中，第一 CEC 或服务器节点 15 被表示成与一个或多个 CEC 或客户机节点连接。

实现本发明的一个例证应用是其中在 CEC 之间保持时钟同步的定时网络。在本发明的说明中，按照服务器时间协议(STP)管理这样的定时网络。按照 STP，在网络中的节点之间发送数据分组，以使每个节点的时钟，例如日时时钟(Time-of-Day clock: TOD 时钟)保持同步。如图 1 中所示，根据 STP，在分组中提供的信息通过耦接链路 16 传递，其中每个 CEC 发送/接收对 STP 网络中的其它系统的请求/响应分组中的数据。在图 1 中描述的定时网络中，可每秒多次在 CEC 15 和每个客户机 CEC 20 之间定期交换定时消息(例如，每 64 毫秒一次)。

假定情况是 CEC 之间的链接遭受临时中断，位于特定 CEC 的时

钟失去同步。既然重要的是了解特定系统的时钟为什么失去与定时网络的剩余部分的同步，因此提出了本发明的系统，以便能够实时收集和特定问题相关的数据。本发明提供一种采取探查器的形式的解决方案，所述解决方案实现同时从多个系统(CEC)收集数据的方法。

如图 1 中所示，根据本发明，提供多个 STP 探查装置 30，每个探查装置通过探查链路 36，与网络中的每个 CEC 节点建立一个路径。每个探查装置 30 被配置成接收归因于在执行 STP 主程序中设置的特殊探查识别异常分支点（probe recognition hook）产生的定时分组。借助主线代码了解探查装置，因为它使用唯一的协同定时网络(CTN)标识符。在 STP 实现中，存在消息间到达时间间隔。探查装置能够检测错误的速度直接和该时间相关。从而，如果消息间时间间隔为毫秒级，那么探查装置能够检测在相同数目的毫秒内的错误。当检查来自 CEC 节点的单个定时分组，或者来自一个或多个节点的分组的组合时，如果探查装置 30 发现问题，那么它产生专用分组，并把专用分组发送给网络中的所有系统，从而把所有相关的调试数据保存到相应的存储装置，例如位于节点的硬盘驱动器，以便稍后分析。由于利用点对点耦接链路的 STP 分组的本质，这种数据收集将在探查装置 30 确定存在错误的很短时间内被触发。一般来说，所述很短时间为微秒级。当在每个系统收到收集相关信息的信号时，在 STP 实现中，存在“冻结”所有相关控制块的状态的选择。从而，当探查装置检测到错误时，它“立即”触发数据收集，而不是等待操作员干预。另外如图 1 中所示，与探查装置相关联的是存储器存储装置 32，用于保存从监视的每个定时分组的字段提取的数据，和记录与在每个节点检测的错误情况相关的信息。

图 2 图解说明根据本发明，由 STP 定时网络中的探查装置实现的方法 50 的流程图。如图 2 中所示，当收到分组时，STP 探查码从在探查系统接收的分组中获得信息，分析所述信息，并立即决定 STP 网络是否在正确工作。在第一步骤 55 中，描述了取回与探查装置连接（attach）的每个 CEC 的标识符的步骤。在该步骤可初始化用于报告

和每个 CEC 有关的信息的日志文件。探查装置动态确定它将监视哪些 CEC。通过专有诊断接口(PDI)，探查装置被启动，从而取回定时分组，分析定时分组以获知是否有异常，并在相连的 CEC 上立即启动数据收集。

从而，如在步骤 58 所示，步骤 58 是进入将实现用于检查每个节点的消息数组，即取回在探查装置接收的每个定时消息并把其内容保存在产生的数组(未示出)中的 PDI 的循环的步骤。如在下一步骤 62，在一个例证的实现中，run_stp_pdi 函数被调用，该函数从发出的分组取回数据(例如一个 256 字节的条目)，并将其保存在为该节点创建的数组中。其数据被提取的定时消息中的特定字段包括(但不限于)：stratum(连接的服务器的层级)；timing_state(指示相连的服务器的定时状态，例如同步、不同步或者时钟被停止)；root_delay 和 root_dispersion(相对于 CEC 时钟源)；timing_mode(相连服务器的定时模式)；和 valid_entry 字段(指示消息是否包含有效的数据)。特别地，作为实现 PDI 函数调用的结果，来自这些消息的数据值被复制到专用于该节点的数组（如果该条目未在先前的 PDI 函数调用中被取回）。除了检查没有条目重复之外，还实现一种特殊的索引，以保证没有任何条目被遗漏。在一个实施例中，在绕回存储区（wrap-round in-storage area）中可保存每个节点 256 个条目。

由于这里描述的例证实施例和定时协议有关，因此探查装置能够通过利用分组中的定时数据，计算它自己和它正探查的 CEC 之间的时钟偏移量。这些计算的偏移量也可被分析，以符合特定的标准。

继续到步骤 65，步骤 65 是处理 PDI 结果的步骤。在一个例证实现中，process_pdi_results 函数被调用，该函数关于触发条件检查数组中的当前消息。根据本发明，这种实时分析允许探查码通知操作员(1)是否已发生“值得注意的事件”。所述“值得注意的事件”是对 STP 网络来说非致命的不寻常事件；或者(2)是否发生了“灾难事件”，响应这种情况，探查装置将向网络中的每个系统发送消息，指令它保存相关数据，并且可能中断执行。

图 3A 图解说明检测触发条件的步骤。在第一实例中，如在步骤 68-69 所示，探查装置将首先检查在消息数据中指示的“定时状态”，并确定是否发生了从同步状态到非同步状态的状态转变。如果是，那么探查装置确定报告的偏移量值保持在容限之内。如果偏移量在容限之外，那么这构成“值得注意的”事件(STP 网络已适当地检测该条件)，并将被照此记录。但是，如在步骤 69 所示，如果确定节点的定时状态已从同步状态转变到非同步状态，并且偏移量保持在容限内，那么这构成“灾难”事件，并将被照此处理。它被认为是灾难性的，因为两个不同的字段值彼此不一致。在灾难情况下，启动另外的调试操作，从而探查装置经由PDI接口产生要求执行被中断并且在STP中的每个节点收集调试数据的控制指令。响应该控制消息，将在 STP 网络的每个相连节点立即记录数据供调试之用。对由探查装置确定的每个灾难事件采取该调试操作。

继续到图 3A 的步骤 70，步骤 70 是确定探查装置计算的时间偏移量是否大于预定值的步骤。图 3B 更详细地描述了步骤 70。在第一步骤 72，确定探查装置计算的在 non-Stratum-1 级服务器的偏移量(SS_offset)和为 Stratum-1 级服务器记录的偏移量(S1_offset)之间的差值的绝对值是否大于 modified_sync_check_threshold 值，例如在 50 和 150 微秒之间，以及其它服务器的定时状态(SS_timing_state)是否同步。即：

|SS_offset-S1_offset|>modified_sync_check_threshold 并且

SS_timing_state = synchronized

如果上述条件成立，那么被认为是灾难事件。

在步骤 72，如果条件不成立，那么处理进入步骤 74，进行第一根离散“完整性检查（sanity check)”，以确定在由除 Stratum-1 级服务器之外的服务器发出的消息中记录的根离散值（root dispersion value），即 SS_root_disp 是否大于 sync_check_threshold 值（sync_check_threshold 值是定时网络的预定值，约为 50 微秒），以及其它服务器的定时状态是否同步。即：

如果 SS_root_disp>sync_check_threshold，并且 SS_timing_state = synchronized

如果上述等式成立，那么再一次，这被处理为灾难事件。

其它触发条件包括对于由正在被分析的协议所允许的预定范围检查字段值。图 3B 的步骤 80 中表示了一个例子，在步骤 80 中，随着它和在服务器的定时消息中关于服务器记录的“Stratum”变化相关，检测触发条件。从而，在步骤 80，关于与在相同服务器的前一定时消息中指示的在先值相比，在最新的定时消息的“stratum”字段中指示的值是否已改变，进行比较。如果 stratum 已改变，那么这表示“值得注意的”事件，并将照此记录操作。另一方面，如果确定在 stratum 字段中指示的值在规定范围之外，它构成“灾难”事件，并将被照此处理。否则，在步骤 80，如果 Stratum 级变化不是灾难性的，那么处理进入图 3B 的步骤 90，检测格式变化触发条件。

在该特定的实施例中，存在必须在定时网络中的所有节点按照协同方式发生的全局事件。触发被定义，以检测何时这些全局变化在所有节点不发生。不同的消息格式和参数关键字（key)（ctn_param_key)一起被用于指示这些全局事件。具体地，如果探查装置确定存在格式变化（借助定时消息的 ctn_param_key 字段)，那么探查装置预期通过检查其它节点各自的定时消息，在规定的时间(即，等于 4 个后续的定时消息的持续时间)内在其它节点上看到相同的变化。如果在四个消息内，对于每个相连的节点，探查装置没有发现格式变化指示(即，定时消息的 ctn_param_key 字段没有变化)，那么这是“灾难性的”事件，并将被照此处理。

从而，根据本发明，进行允许探查码把几种事件通知操作员的实时分析。“值得注意的事件”是对 STP 网络来说非致命的不寻常事件。这种情况下，探查码将把该事件连同日期和时间一起记录到探查装置的日志中。值得注意的事件的一个例子是当在先分组有效时，无效的定时分组的接收，或者反之亦然。探查装置还检测“灾难性事件”。灾难性事件的例子包括(但不限于)：(a)分组指示不在所支持的范围中的

stratum 层级；(b)在可接受的时间范围内没有收到所需的格式变化，并且没有分组被遗漏；(c)收到指示定时网络不同步的分组，但是时间偏移量指示网络仍然在可接受的容限内；以及(d)在格式方面变化，而不是关键字方面的变化。

一旦探查装置 30 确定发生了"灾难性事件"，那么探查装置通过探查链路向网络中的每个系统发送一个消息，指令该系统保存相关数据，并且可能中断执行。这种处理将保证在网络中的每个系统获得用于所检测的问题的调试数据，并将为数据分析和问题确定及解决创造条件。

在本实施例中，探查装置还可被扩展到检测两个时钟之间的长期时钟漂移率。例如，可通过利用独立的专用探查链路，经由来自一个或多个探查机器 30 的探查，监视在图 1 中的各个节点，即设备 15、20 的时钟。每个探查机器向所有客户机和服务器发送探查分组。探查机器并不调整它们的本地机器时钟。与通过观察延迟测量结果来检测时钟偏移量和偏斜类似，这些延迟测量结果揭示被探查的设备上的时钟和探查机器上的时钟之间的相对时钟行为。这种方法使得能够观察所有其它时钟相对于相同本地时钟的行为。与相同时钟在相对行为方面的差异显示被探查的这些时钟之间的差异。事实上，使用 GPS(全球定位系统)或 PPS(每秒脉冲数服务)信号的全球时钟正在使用相同的抽象概念，因为 GPS 或 PPS 信号起探查站的作用，并且在本地时钟级计算时钟和信号之间的差异。

从而，在一个实施例中，探查可被用于检测两个时钟之间的长期时移漂移。探查还可被用于检测可能在客户机和服务器之间，或者在不同的客户机之间，或者在运行不同的同步算法的客户机之间的多个时钟方面的差异。探查还可被用于研究在各种不利的环境（例如拥塞的链路，断开的链路(down link)等）下，不同的同步算法的行为。在一个例证实现中，在建立同步环境（包括建立探查机器 30 并使它们与服务器和客户机相连）之后，每隔一定时间从探查机器向服务器和客户机传送探查分组。从探查分组获得的数据可被分析，从而获得被探

查的机器的动作。例如，这样的数据可被用于描绘来自被探查的每个机器的前向延迟和负后向延迟；获得前向延迟和负后向延迟之间的中心线，以描述被探查的机器的时钟行为；或者描绘来自被探查的所有机器的中心线并进行比较。

上面参考根据本发明的实施例的方法、设备(系统)和计算机程序产品的图表，说明了本发明。显然每个图表可由计算机程序指令实现。这些计算机程序指令可被提供给通用计算机，专用计算机，嵌入式处理器或其它可编程数据处理设备的处理器，从而产生一台机器，以致通过计算机或者其它可编程数据处理设备的处理器执行的指令产生用于实现这里规定的功能的装置。

这些计算机程序指令还可被保存在能够指令计算机或其它可编程数据处理设备按照特定的方式起作用的计算机可读存储器中，以致保存在计算机可读存储器中的指令产生包括实现这里规定的功能的指令装置的制造产品。

计算机程序指令还可被装入计算机可读的或者其它可编程数据处理设备，导致在计算机或其它可编程设备上执行一系列的操作步骤，从而产生计算机实现的进程，以致在计算机或其它可编程设备上执行的指令提供用于实现这里规定的功能的步骤。

应明白本发明可用在实现其它连网协议，从而在计算机设备的互连网络的节点间传递信息，以便调试这样的节点的系统中。虽然显然这里公开的发明非常适合于实现上面陈述的目的，不过要认识到本领域的技术人员可设计出各种修改和实施例，附加的权利要求覆盖落入本发明的精神和范围内的所有这样的修改和实施例。

图1

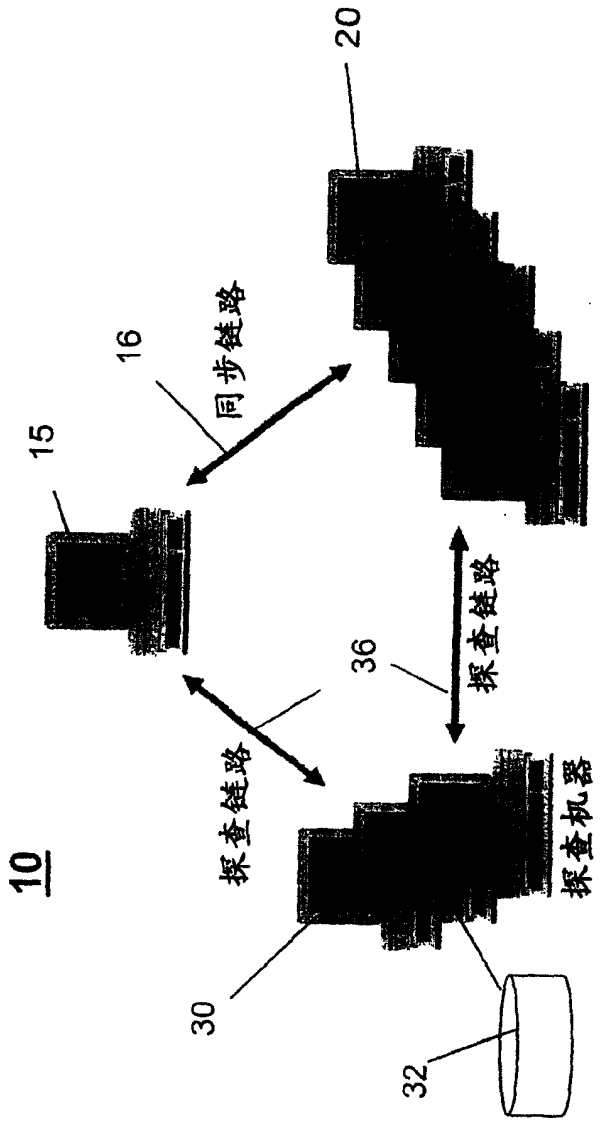


图 2

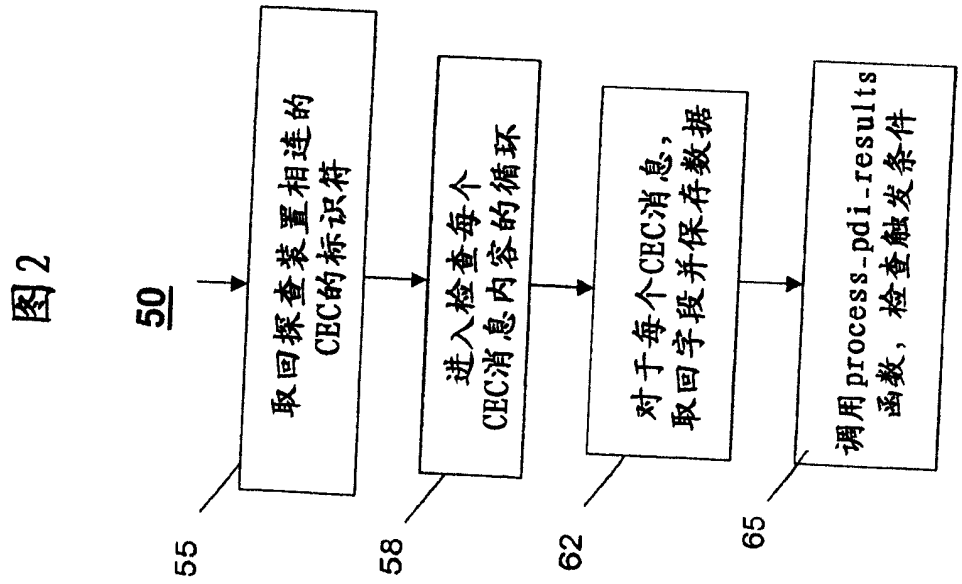
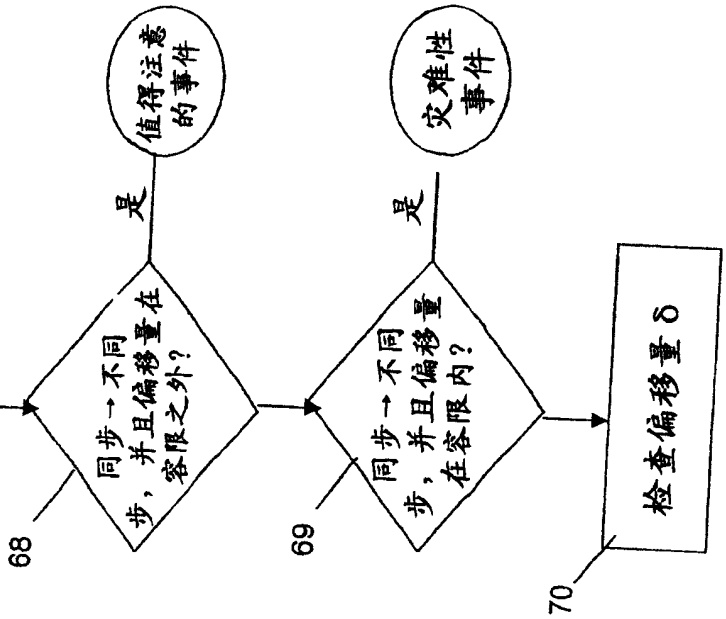

50
取回探查装置相连的CEC的标识符
55
进入检查每个CEC消息内容的循环
58
对于每个CEC消息，取回字段并保存数据
62
调用process-pdi-results函数，检查触发条件
65

图 3A


65
同步→不同步，并且偏移量在容限之外？
68
是
值得注意的事件
同步→不同步，并且偏移量在容限内？
69
是
灾难性事件
检查偏移量 δ
70

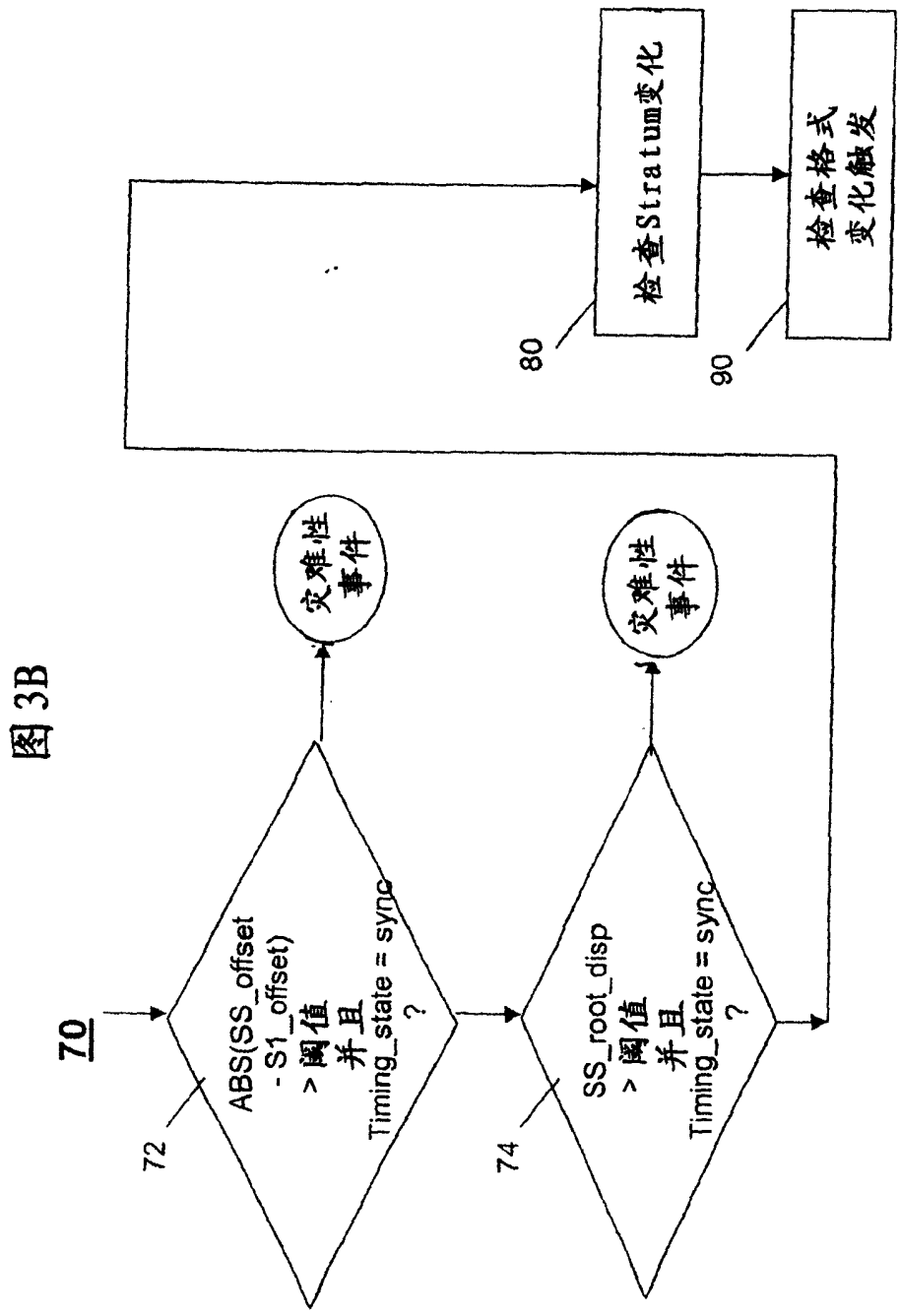

70
72
ABS(SS_offset
- S1_offset)
> 阈值
并且
Timing_state = sync
?
灾难性
事件
74
SS_root_disp
> 阈值
并且
Timing_state = sync
?
灾难性
事件
80
检查Stratum变化
90
检查格式
变化触发

图 3B