发明名称
使固件能从 iSCSI 设备引导系统的方法、系统和设备

摘要
一种使系统固件能够从网络连接的因特网小型计算机系统接口 (iSCSI) 设备高效地引导操作系统 (OS) 和 / 或客户端程序的方法和系统。该方法一般包括：(1) 定义 iSCSI 设备在表示系统硬件的分级数据结构内的固件表示，以及 (2) 扩展网络支持包以容纳额外的引导变量。该额外的引导变量允许系统固件获取从该网络连接的 iSCSI 设备进行引导所需的信息，而同时使用现有发现协议之一。
1. 一种用于完成并扩展网络存储包的方法，包括：
从客户端系统的系统文件接收引导变元组；
确定所述引导变元组是否包含命令关键词；
当在所述引导变元组中存在命令关键词时，自动激活因特网小型计算机系统接口
iSCSI 设备发现过程，其中 iSCSI 设备被访问以检索用于所述客户端系统的引导文件；以及
当不存在所述命令关键词时，启动标准的简单文件传输协议 TFTP 自举过程。
2. 根据权利要求 1 所述的方法，进一步包括：
确定所述引导变元组是否包含一个或多个限定符关键词；
当一个或多个限定符关键词位于所述引导变元组内时，所述自动激活包括：
启动分别对应于所述一个或多个限定符关键词的一个或多个特定发现过程，其中所述
一个或多个特定发现过程访问特定信息服务器以检索寻址参数和用于所述 iSCSI 设备的
其它访问信息；以及
当从所述特定信息服务器接收到所述寻址参数和其它访问信息时，将所述寻址参数和
其它访问信息转发给所述 iSCSI 设备的盘支持包，以触发将所述引导文件从所述 iSCSI 设
备加载到所述客户端系统。
3. 根据权利要求 2 所述的方法，其中所述启动特定发现过程启动与包含在所述引导变
元组内的所述一个或多个限定符关键词相关的多个发现过程，其中所述多个发现过程返
回所述寻址参数和其它访问信息中所述的找寻地址参数，所述限定参数被合并以提供一组完
整的所述需寻址参数和其它访问信息；
其中，所述返回所述引导文件给所述客户端系统是由所述 iSCSI 设备中的所述盘支持包在
接收到所述一组完整的所述需寻址参数和其它访问信息时来完成。
4. 根据权利要求 3 所述的方法，其中当从不同信息服务器检索多个相同参数时，所述
网络支持包完成选择从最后查询的服务器中检索的参数值并用以在所述一组完整的所述需
寻址参数和其它访问信息中使用的能力。
5. 根据权利要求 3 所述的方法，其中所述命令关键词是 iscsi，并且所述一个或多个限
定符关键词包括 dhcp，从而所述 iSCSI 设备的寻址参数是从动态主机配置协议服务器检索
的。
6. 根据权利要求 2 所述的方法，进一步包括：
在所述硬盘生成的已扩展引导变元组内包括第一变元和第二变元，其中所述第一变元
是所述命令关键词，其区别不同类型的自举方法，而所述第二变元是所述一个或多个限定
符关键词，其指定发现协议用以获取访问所述 iSCSI 设备和从所述 iSCSI 设备加载所述引
导文件所需的的地址参数；
在所述引导变元内接收网络专用引导参数，其指示所述 iSCSI 设备被表示为由所述客
户端系统的固件生成的分级数据结构内的类网络的设备类型，其中所述网络支持包处理
所述网络专用引导参数并经由发现过程获取所述 iSCSI 设备的另外的地址参数和访问信
息，以使能 (a) 当 iSCSI TCP 减载引擎 TOE 硬件不存在时经由软件 iSCSI 启动器建立 iSCSI
会话，以及 (b) 经由所述 iSCSI 会话从所述 iSCSI 设备检索所述引导文件。
7. 一种用于使固件能从 iSCSI 设备引导系统的方法，所述方法应用在具有客户端系
统、因特网小型计算机系统接口 iSCSI 设备和一个或多个信息服务器的传输控制协议 / 因
特网协议计算机网络环境中，该方法包括：

在所述客户端系统的固件的初始引导过程中，在所述客户端系统的硬件配置的分级
数据结构内表示所述 iSCSI 设备；

生成传输给网络支持包的已扩展引导变元组，其中所述已扩展引导变元包括命令关
键词，所述命令关键词向所述网络支持包标识出用于从所述 iSCSI 设备检索引导映像的
iSCSI 发现过程；以及

将所述已扩展引导变元组转发到所述网络支持包，用以触发所述网络支持包实现所述
iSCSI 发现过程。

8. 根据权利要求 7 所述的方法，进一步包括：

于所述已扩展引导变元内添加一个或多个限定符关键词，其向所述网络支持包标识出
特定发现协议，用以实现从所述一个或多个信息服务器检索所述 iSCSI 设备的地址参数和
其它访问信息。

9. 根据权利要求 8 所述的方法，其中进一步地，所述一个或多个限定符关键词进一步
向所述网络支持包标识出将访问所述一个或多个信息服务器中的哪一个，用来检索访问所
述 iSCSI 设备以及从所述 iSCSI 设备加载所述引导映像所需地址参数和其它信息。

10. 根据权利要求 8 所述的方法，其中所述一个或多个限定符关键词之一标识特定的
发现协议和使用所述特定的发现协议所需的另外的引导变元。

11. 根据权利要求 8 所述的方法，其中在所述已扩展引导变元组内提供多个限定符关
键词时，所述多个限定符关键词标识各发现协议和相关联的信息服务器，从而所述网
络支持包可完成多个发现过程，其集中地检索访问所述 iSCSI 设备以及加载所述引导映像
所需的地址参数和其它访问信息。

12. 根据权利要求 8 所述的方法，其中所述命令关键词和所述一个或多个限定符关键
词被包括在所述已扩展引导变元组内。

13. 根据权利要求 7 所述的方法，其中所述命令关键词是互斥的多个可用命令关键词
中的一个命令关键词，从而在所述已扩展引导变元组中只可提供一个命令关键词。

14. 根据权利要求 7 所述的方法，其中

提供所述命令关键词用于 iSCSI 设备和用于网络设备，从而在所述客户端系统内不
存在任何 iSCSI TCP 减载引擎硬件时命令关键词触发对 iSCSI 启动器的使用以建立 iSCSI
会话；以及

当所述 iSCSI 设备在所述分级数据结构中被表征为类网络的设备类型时，所述方法进
一步包括：

提供网络专用引导参数给所述网络支持包，其中所述网络支持包 处理所述网络专用
引导参数并且经由发现过程获取所述 iSCSI 设备的另外的地址参数和访问信息，以使能从
所述 iSCSI 设备检索引导映像；以及

当 iSCSI TCP 减载引擎硬件不存在时通过提供所述 iSCSI 启动器来建立 iSCSI 会话。

15. 根据权利要求 7 所述的方法，进一步包括：

在所述固件生成的所述已扩展引导变元组内包括第一变元和第二变元；

其中所述第一变元包括命令关键词，其对应者特定自举方法和相关联的引导变元，并
且区别不同类型的可用自举方法；

3
其中叙述第二变元包括一个或多个限定符关键词，其定义发现协议用以获取访问所述iSCSI设备以及从所述iSCSI设备加载引导映像所需的地址参数；以及

变元在所述网络支持包内实现一种机制，其基于在所述引导变元组内的命令关键词的存在，区分不同类型的自举方法，所述自举方法包括TFTP自举和iSCSI自举；

其中，所述TFTP自举是缺省方法，从而在所述已扩展引导变元组内缺少命令关键词时触发TFTP引导方法。

16. 一种使固件能从iSCSI设备引导系统的系统，包括；

具有固件的客户端设备，所述固件通过生成分级数据结构来启动所述客户端设备的引导过程，所述分级数据结构包括对iSCSI设备的表示；

用于经由所述固件生成用来从iSCSI设备检索引导文件的已扩展引导变元的装置，其中用于生成已扩展引导变元的所述装置包括用于在所述引导变元内提供命令关键词和一个或多个限定符关键词的装置；

网络支持包，其被配置成接收和处理包括所述命令关键词和所述一个或多个限定符关键词的所述已扩展引导变元，目的从特定信息服务器获取地址参数和其它访问信息，所述特定信息服务器的发现协议是由所述已扩展引导变元内的所述限定符关键词来标识；以及

用于从所述iSCSI设备检索所述引导文件并将所述引导文件转发到所述客户端设备以便完成所述引导过程的装置。

17. 根据权利要求16所述的系统，其中所述网络是传输控制协议/因特网协议计算机网络，其具有（a）客户端设备，（b）iSCSI设备，所述iSCSI设备在其上具有用于完成引导所述客户端设备的引导文件并位于物理网络上，以及（c）可经由对应的发现协议来访问的一个或多个信息服务器，并且所述一个或多个信息服务器提供寻址信息以使能对所述iSCSI设备的定位和访问，所述系统包括用于以下功能的装置；

在所述客户端设备的固件初始引导过程期间，在所述客户端设备的硬件配置中的分级数据结构内表示所述iSCSI设备；

生成用于传输给网络支持包的一组已扩展引导变元，其中所述已扩展引导变元包括命令关键词，其向所述网络支持包标识出需要用于从所述iSCSI设备检索引导文件的iSCSI发现过程；

在所述已扩展引导变元内添加一个或多个限定符关键词，其向所述网络支持包标识出特定发现协议，用以实现从所述一个或多个信息服务器检索所述iSCSI设备的地址参数和其它访问信息；以及

将所述一组已扩展引导变元转发给所述网络支持包以触发所述网络支持包实现所述iSCSI发现过程。

18. 根据权利要求17所述的系统，其中进一步地；

所述命令关键词和所述一个或多个限定符关键词在所述一组已扩展引导变元内；

所述命令关键词是互斥的多个可用命令关键词中的一个命令关键词，从而在所述一组已扩展引导变元内可以提供一个命令关键词；

所述一个或多个限定符关键词进一步向所述网络支持包标识出将访问所述特定信息服务器中的哪一个，用于检索访问所述iSCSI设备以及从所述iSCSI设备加载引导文件所
需的所述地址参数和其它访问信息；

至少一个限定符关键词标识特定义发现协议以及使用所述特定义发现协议所需的另外的引导变量；以及

当在所述一组已扩展引导变量内提供了多个限定符关键词时，所述多个限定符关键词标识出各个发现协议和相关联的信息服务器，从而所述网络支持包可完成多个发现过程，其集中地检索访问所述 iSCSI 设备并加载所述引导文件所需的所述地址参数和其它访问信息。

19. 根据权利要求 16 所述的系统，其中进一步地；

提供所述命令关键词用于 iSCSI 设备和用于网络设备，从而当在所述客户端设备内不存在任何 iSCSI TCP 减载引擎硬件时所述命令关键词触发对 iSCSI 启动器的使用以建立 iSCSI 会话；以及

当所述 iSCSI 设备在所述分级数据结构中被表征为类网络的设备类型时，所述系统进一步包括用于以下功能的装置；

提供网络专用引导参数给所述网络支持包，其中所述网络支持包处理所述网络专用引导参数并且经由发现过程获取所述 iSCSI 设备的另外的地址参数和访问信息，以使能从所述 iSCSI 设备检索引导文件；以及

当 iSCSI TCP 减载引擎硬件不存在时通过提供所述 iSCSI 启动器来建立 iSCSI 会话。

20. 根据权利要求 17 所述的系统，进一步包括用于以下功能的装置；

扩展网络支持包以在由所述固件生成的所述一组已扩展引导变量内容纳第一变元和一个或多个第二变元；

其中所述第一变元是命令关键词，其对应着特定义举方法和相关联的引导变元，并且区别不同类型的可用自举方法；

其中所述一个或多个第二变元是一个或多个限定符关键词，其指定发现协议用于获取用于访问所述 iSCSI 设备以及从所述 iSCSI 设备加载所述引导文件的另外的地址参数；以及

在所述网络支持包内建立一种机制，其基于所述一组已扩展引导变量内的所述命令关键词的存在，从 TFTP 自举和 iSCSI 自举中区别不同类型的自举方法；

其中所述 TFTP 自举是缺省方法，从而在所述一组已扩展引导变量内 缺少所述命令关键词触发所述 TFTP 自举方法。
使固件能从 iSCSI 设备引导系统的方法、系统和设备

技术领域
[0001] 本发明一般地涉及计算机系统，并且具体地涉及计算机系统的引导过程，以及在计算机系统中使用网络支持包（network support package）。具体地，本发明涉及一种用于从因特网小型计算机系统接口（iSCSI）设备完成计算机系统的引导过程的方法和系统，用于计算机系统的远程网络引导操作（boot operation），以及用于使用网络支持包从引导服务器加载计算机系统的引导映像的增强方法。

背景技术
[0002] 紧随着计算机系统的上电（或重新引导），系统经历引导过程；借助该引导过程，系统的固件通过标识以及设置/启动系统中的设备而准备该系统以便进行操作。在引导过程中，固件将系统硬件表示（represent）成带有附属设备的互连总线的分层体系。这种分层表示通常描述用户配置，并且包含用于硬件设备的固件设备驱动程序和供那些驱动程序使用的支持例程。

[0003] 另外，固件通过从某个存储位置/设备加载（引导）操作系统来准备该计算机系统以便进行操作。在单机计算机系统的正常工作期间，当用户向计算机发出引导命令时，计算机响应于该引导命令，尝试从该计算机系统的本地存储器检索操作系统文件。还需要配置数据文件来用特定硬件配置所必须的硬件参数来配置特定系统。

[0004] 某些计算机系统经由网络连接到服务器或网络可访问的其它设备。在使用单机系统（其中当关断电源时在本地维护计算机的存储装置）时，OS 典型地存储在计算机系统自身中。在具有当关断电源时会丢失的临时（易失性）存储装置的其它系统中（或者甚至在使用具有有限的非易失性存储能力的系统时），计算机不能从该计算机自身内检索引导信息。在这种情况下，计算机系统的固件可配置成经由网络向用作引导服务器的服务器发送针对 OS 文件的请求。远程引导操作在分布式和网络计算的世界中变得常见，特别是在具备从引导服务器进行引导的优势的情况下，诸如节省计算机系统的有限的存储器资源。

[0005] 当创造出新设备时，现有计算机系统典型地不能从这些新设备引导客户端程序或操作系统，除非固件提供对该引导过程的某种类型的支持。当新设备不适合预先定义的设备类型或系统中典型地存在的标准设备类型时，现有系统（在缺少特别提供的固件的情况下）不具备完成针对新设备的引导过程的这种能力是尤其正确的。正被更加频繁地使用但尚未被提供直接固件支持的一种这样的新设备是因特网小型计算机系统接口（iSCSI）设备。

[0006] iSCSI 协议允许客户端系统使用 iSCSI 启动器来连接诸如 IP 网络上的磁盘或磁带驱动器之类的远程目标以进行块级输入/输出（I/O）。SCSI 体系结构是基于客户端/服务器模型，其中客户端（“启动器”）典型地是发出请求以读取写入的诸如文件服务器之类的主机系统，而服务器（“目标”）是对客户端请求做出响应的诸如磁盘阵列之类的资源。服务器（目标）包括一个或多个被分配有标识号或逻辑单元号（LUN）的逻辑单元。本领域的技术人员熟知 iSCSI 协议和它的一般实现。在 IETF（因特网工程任务组）的 RFC 储
存库中公布的 RFC 3720:Internet Small Computer Systems Interface (iSCSI) 中可获得该协议的更多描述。

[0007] 因为 iSCSI 协议还未被完全集成进诸如由固件指导的远程引导操作之类的现有系统功能，所以常规系统使用若干现有发现协议以从公知服务器获得关于 iSCSI 客户端和目标的信息。由 IETF RFC 4173 在网站“www.ietf.org/rfc/rfc4173.txt”上在线发表的标题为“Bootstrapping Clients using the iSCSI Protocol”的文章中提供了对这些现有发现协议的描述。然而，当前没有任何高效方法，能够提供客户端通过 iSCSI 设备（目标）引导操作系统的受指导的引导过程。因此，本发明认识到存在以下需求：由 iSCSI 客户端从 iSCSI 设备（目标）高效地引导其的操作系统，而无需完全重新设计/修改现有的远程引导过程和发现协议。

[0008] 网络计算机系统在计算机领域中是众所周知的。利用普通文件传送协议 (TFTP) 从位于跨网络的远程服务器加载计算机系统的引导映像/文件的通用方法也是已知的。特别地，当在诸如局域网 (LAN) 的网络上引导计算机系统（即，网络设备）时，计算机系统的固件通常利用 TFTP 自举 (bootstrap) 方法来从跨网络的引导服务器/设备获得引导文件。

[0009] 利用该 TFTP 自举方法，提供了网络支持包来辅助网络引导操作。具体而言，网络支持包利用 TFTP 自举方法来检索和/或定义访问和检索位于引导服务器处的引导映像/文件所需要的特定引导变量 (boot arguments)。网络支持包还处理从计算机系统获得的引导变量，并建立与远程服务器的连接来获得引导文件。可以在开放固件工作组 (Open Firmware Working Group) 发布的开放固件推荐规范 (Open Firmware Recommended Practice; TFTP 引导扩展版本 1.0 (TFTP BootingExtension Version 1.0)) 中找到关于利用网络支持包来启用由经由 TFTP 自举位于跨 LAN 的引导服务器加载引导映像的概要信息。

[0010] 虽然以上 TFTP 自举方法被应用于基于 LAN 的和其它类似配置的网络内并且运行良好，但是已经证明这样的方法在不同的网络配置或网络类型的情况下并不有效。举例来说，TFTP 自举方法在 InfiniBand (IB, 无限带宽) 网络的情况下无效。

[0011] 如本领域中已知的，InfiniBand (IB) 网络是一种分组交换的窄并行 (narrow parallel) 网络，其操作于高数据速率和大物理范围，从而将远程设备互连至主计算机系统（或服务器）。IB 网络包括将主计算机系统耦合至子网的主机通道适配器 (HCA)。该网络进一步包括一个或多个目标通道适配器 (TCA)，而这些一个或多多个目标通道适配器 (TCA) 又将其它设备耦合至子网。子网通常包括至少一个交换机以及将 HCA 和 TCA 连接到交换机的多条链路。例如，简单的 IB 网络可以具有一个交换机，HCA 和 TCA 通过相应的链路连接至该交换机。

[0012] IB 网络常常分组适配器和路由器连接至局域或广域通信网络。例如，可以利用以太网适配器来传输在以太网上的通信。网络适配器具有其自己的 TCA，作为 IB 网络内用于耦合于以太网的端节点。IB 规范进一步提供了一种原始数据报通信模式，其接收到以太网接收到的分组用于在 IB 网络上传输，并且反之亦然。类似地，IB 网络可以包括连接至因特网的路由器或适配器。

[0013] 假定提高对完成计算机设备的引导的远程引导操作的信赖，并且诸如 IB 网络这样的新网络的复杂性增加，那么需要扩展在 LAN 内完成的远程引导过程，以便使得主计算机能够完成来自位于跨 IB 网络的引导服务器的远程引导过程。目前还没有支持经由连接
至 IB 结构的网络设备的标准网络引导过程来无缝实现网络引导操作的开放固件，其中该引导设备直接连接至 IB 结构或连接至耦合于 IB 结构的端节点的外部网络。

[0014] 网络计算机系统通常在本领域中是公知的。通过网络远程引导服务器加载计算机系统的引导映像的常规方法也是公知的。当通过网络引导计算机系统时，计算机系统的固件通常使用简单文件传输协议（TFTP）通过网络从引导设备获取引导文件。提供了网络支持包通过处理从计算机系统获取的引导变量以及建立与远程服务器的连接来帮助此远程（网络）引导操作获取引导文件。网络支持包使用 TFTP 自举方法来检索和 / 或定义特定的引导变量。有关使用网络支持包以使能从远程引导服务器加载引导映像的一般信息可以在 Open Firmware Recommended Practice: TFTP Booting Extension Version 1.0（由 Open Firmware Working Group 发布）中找到。

[0015] 标准网络支持包使用 TFTP 自举方法以通过网络引导系统。TFTP 自举方法的引导变量提供了指定自举协议（BOOTP）协议以查找引导服务器和引导文件的选项。因此，引导变量的当前（现有）定义提供了例如使用 BOOTP 而不是 RARP/ARP（地址解析协议）以查找引导服务器并获取引导文件的选项。如果未指定 BOOTP 协议，则可以使用诸如 RARP/ARP 之类的其他协议获取引导文件。此外，使用常规方法，以 IPv4 格式（网际协议版本 4）指定所有表示网络地址的变量。

[0016] 不断的发展已导致创建更高级的 IP 格式，例如 IPv6。但是，用于启用远程引导操作的现有变量和相应方法仍使用 IPv4 格式，因为先前尚未定义 IPv6 格式。此外，使用现有方法，尚未将动态主机配置协议（DHCP）视为用于 TFTP 自举的必选引导发现协议。因此，本发明认识到需要升级远程引导操作的现有方法以便允许使用 IPv6 格式和 DHCP 获取使用远程服务器上存储的引导文件来引导系统所需的信息。

发明内容

[0017] 公开了一种使系统固件能够从网络连接的因特网小型计算机系统接口（iSCSI）设备高效地引导操作系统（OS）和 / 或客户端程序的方法和系统。该方法一般包括：(1) 定义 iSCSI 设备在表示系统硬件的分级数据结构内的固件表示；以及 (2) 扩展网络支持包以容纳额外的引导变量，该额外的引导变量允许系统固件获取从该网络连接的 iSCSI 设备进行引导所需的信息，而同时使用现有发现协议之一。

[0018] 从理解中提供了本发明的特定方法：为了从 iSCSI 设备引导系统，该系统的引导固件需要识别 iSCSI 硬件或使能 iSCSI 软件启动器、标识 iSCSI 客户端和目标 iSCSI 设备，与该目标建立通信，以及从该目标加载引导文件。表示设备的固件数据结构的这种分级性质允许信息在整个设备和子设备之间以及在设备与支持包之间进行传递。根据一个实施例，iSCSI 设备类型被表征为类网络的设备类型，目的是支持 leverage 已扩展的网络支持包并因此使用网络服务来发现和获取 iSCSI 设备地址的不同分量。通过这种分级表示，iSCSI 设备提供网络专用引导参数（例如速度、模式等等）并将那些引导参数传递给网络支持包。网络支持包获取并处理 iSCSI 设备地址中的分量并将那些变量（也即引导参数）传递给盘支持包（disk support package）。接着，盘支持包加载引导映像并将 OS 转发给 iSCSI 客户端以使能在 iSCSI 客户端处完成对 OS 的引导。

[0019] 公开了一种方法和系统，用于使得 InfiniBand（IB）主计算机设备能够通过 IB 网
络从远程引导设备引导操作系统。增强所述设备的固件，以便在表示计算机设备的硬件的层次数据结构内提供 IB 主机通道适配器 (HCA) 的表示。相应地，定义新的设备类型，“hca”设备类型，其表示 IB HCA。hca 设备在层次固件数据结构内通过节点全局唯一标识符 (node-GUID) 来描述特征。所述 hca 设备包含用于开启和关闭设备的接口以及从设备读取和写入设备的接口。所述 “hca”设备包括多个“子设备”，其被称为 “ib 端口 (ipbport)” 设备，并且由端口全局唯一标识符 (port-GUID) 来标识。除了为父 hca 设备定义的开启、关闭、读和写接口以外，ib 端口设备还包括使得该设备能够被用作引导设备的加载接口。

创建 “ib- 引导 (ib-boot)” 支持包来协助封装用于使用 ib 端口设备通过 IB 网络进行引导的变元，并且定义引导变元集来由所述 ib- 引导支持包利用。所述 ib- 引导支持包对指定目标 IB 端节点以及一个或多个关键字的引导变元进行处理。在一个实施例中，所述 ib- 引导支持包还实现特定的分组传送协议，当通过 IB 网络进行引导时，通常通过 hca 设备驱动器来利用所述特定的分组传送协议。

利用固件结构的层次特性，基于指定关键字来将这些 IB 专用变元 (IB-specific arguments) 传递至适当的支持包，以便完成引导处理。在一个实施例中，提供了两个关键字，各自标识支持包和套接字直接协议 (Sockets Direct Protocol, SDP) 支持包中特定的一个。第一关键字指定要实现 IPoIB (基于 IB 的网际协议 (IP)) 网络引导方法，由此使用 IPoIB 网络协议来到达引导服务器并且找到引导文件。第二关键字指定要实现 IB 网络引导方法，由此使用 SDP 网络协议来访问引导设备。因而通过支持包之一提供了对所述引导服务器的访问，并且经由 IB 网络将引导映像返回给主机设备用于完成引导操作。

披露了一种扩展网络支持包的功能以接受 IPv6 格式的网络地址并允许使用诸如动态主机配置协议 (DHCP) 之类的其他引导发现协议以便获取使用来自远程引导服务器的引导文件引导系统所需的信息的方法和系统。向所述网络支持包添加了一种机制以便能 / 允许支持其他引导发现协议和其他网络地址格式。所述机制定义了添加到引导变元并由所述网络支持包处理的限定符关键字。所述限定符关键字是由固定文件变元 (即，特定的引导发现协议)，变元格式以及与变元关联的其他参数的方法来修改与自举方法关联的引导变元。可选地在与所述自举方法关联的变元之前指定限定符关键字以便修改引导变元的处理。一次可以向引导变元添加多个限定符关键字，并且可以按照关键字在所述变元中的顺序来应用所述关键字。

在一个实施例中，定义了两个限定符关键字，即 ipv6 和 dhcp。可选地将 ipv6 关键字添加在第一引导变元之前以指示将使用 IPv6 格式来表达引导变元中提供的任何网络地址。没有关键字“ipv6”指示以 IPv4 格式表达所有网络地址。dhcp 关键字用于指示网络支持包应使用 DHCP 服务器来获取引导参数。可以可选地提供 diaddr 值以指定要使用的 DHCP 服务器的因特网地址。否则，所述网络支持包执行广播查询以查找 DHCP 服务器。diaddr 值以 IPv6 或 IPv4 表示法表达，具体取决于是否指定了可选的限定符关键字 ipv6。

公开了一种方法和系统，用于使得 InfiniBand (IB) 主计算机设备能够通过 IB 网络从远程存储设备引导操作系统。增强所述设备的固件，以便在表示计算机设备的硬件的层次数据结构内提供 IB 主机通道适配器 (HCA) 的表示。相应地，定义新的设备类型，“hca”设备类型，其表示 IB HCA。hca 设备在层次固件数据结构内通过节点全局唯一标识符来描述特征。所述 hca 设备含有用于开启和关闭设备的接口以及从设备读取和写入设备的接口。
具体实施方式

本发明提供一种使系统固件能够从网络连接的因特网小型计算机系统接口（iSCSI）设备有效地引导操作系统（OS）和/或客户端程序的方法和系统。该方法一般包括：（1）定义iSCSI设备在表示系统硬件的分层数据结构内的固件表示；以及（2）扩展网络支持包以容纳额外的引导变元，该额外的引导变元允许系统固件获取从该网络连接的iSCSI设备进行引导所需的信息，而同时使用现有发现协议之一。

在以下对本发明的示例实施例的详细描述中，足够详细地描述了其中可实现本发明的具体示例实施例，以使本领域的技术人员能够实现本发明，并且应当理解，还可使用其它实施例，以及在不偏离本发明的精神或范围的条件下可做出逻辑的、体系结构的、程序的、机械的、电气的以及其它的改变。由此，以下详细描述不是理解为限制的意思，并且本发明的范围仅由所附各权利要求限定。
在附图的描述中，类似的组件被提供与先前图中的那些组件相类似的名字和标号。当后面的图在不同环境中使用该组件或使用具有不同功能的该组件时，该组件被提供表示该图号的不同前导数字（例如 11xx 用于图 1A 以及 12xx 用于图 1B）。分配给组件的具体数字仅被提供用以帮助进行描述并且其意思不是暗示对本发明的任何限制（结构的或功能的）。

现在参考附图，图 1A 是图示对由网络 1160 连接到远程 iSCSI 设备 (1170)（例如，网络存储装置、引导服务器，或目标）的计算机系统 1100 以及一个或多个信息源 1180 的结构性表示的框图。贯穿对说明性实施例的该描述，iSCSI 设备 1170 可以是任何称为“目标”或者“iSCSI 服务器”或者“引导服务器”或者“客户端”或者“服务器端”或者“iSCSI 客户端”的计算机系统 1100 可以是个人计算机、服务器，移动或手持电子设备，或者其它数据处理设备。网络 1160 可以是局域网、城域网、广域网，或其它类型的计算机网络。网络 1160 可以是有线或无线并且可使用诸如因特网协议（IP）之类的无连接的分组交换协议和/or 其它协议。

计算机系统 1100 包括将计算机系统 1100 连接到网络 1160 的网络接口设备 1110。网络接口设备 1110 表示计算机系统 1100 的硬件部件并且可以是网络接口卡（NIC）或集成的网络设备（例如构建到计算机系统 1100 中的诸如母板之类的主电路板的网络适配器）。

设备驱动器 1120 控制网络接口设备 1110，提供一种管道，通过所述管道操作系统 (OS) 1125 以及计算机系统 1100 的应用层中的一个或多个软件应用 1150 可以与网络 1160 进行通信。设备驱动器 1120 允许通过使用网络接口设备 1110 通过网络 1160 来执行数据存储操作 1130。例如，设备驱动器 1120 可包括 iSCSI 启动器，其允许 SCSI 命令在网络 1160 上与目标目标 (iSCSI 设备 1170) 相互作用。

设备驱动器 1120 中的至少一个可以支持使用网络 1160 引导 iSCSI 设备 1170，在这种情况下 iSCSI 设备 1170 表示远程引导设备。作为远程引导设备，iSCSI 设备 1170 在经由网络 1160 完成的一系列过程将（操作系统的）引导映像加载到计算机系统 1100 之前被初始化。iSCSI 设备 1170 可位于靠近或者远离计算机系统 1100 的物理空间。例如，iSCSI 设备 1170 可通过局域网 1160（例如，吉比特以太网）连接到计算机系统 1100 并且可与计算机系统 1100 位于相同的建筑，机架或机框中。

另外，设备驱动器 1120 允许使用网络接口设备 1110 通过网络 1160 来执行网络任务操作 1140。设备驱动器 1120 中的至少一个可以向操作系统将自身呈现为网络设备驱动器并支持通用网络服务。因此，操作系统可查看并与驱动器 1120 之一交互作用，仿佛该驱动器是在机器引导过程中的总线驱动器和存储装置驱动器之后被加载的常规网络设备驱动器。所支持的网络服务取决于目的可以包括多个不同的组网协议，诸如 TCP/IP、ODI（开放数据链路接口）、PXE（预引导执行环境）、以及各种其它经标准化的和/or 基于公司/or 操作系统的组网协议。通常，设备驱动器 1120 允许计算机系统 1100 使用多个组网协议与一个或多个信息源 1180 进行通信。

虽然计算机系统 1100 的设备驱动器 1120 在图 1A 中呈现在与操作系统和硬件相分开的层中，但是应当理解，取决于实际设备驱动器可在计算机系统 1100 内的不同环境中操作。从断电、经过引导序列、直到到达其中操作系统被完全加载并控制了计算机系统 1100 的操作模式的模式操作。此外，可以用硬件、固件和/or 软件来实现此处所描述的各种驱动
器功能。

[0058] 如上所述，设备驱动器 1120 可支持对远程引导设备 (iSCSI 设备 1170) 的引导并支持通用网络业务，而设备驱动器 1120 中的至少一个可以向操作系统将自身呈现为网络设备驱动器。可以通过使用 iSCSI 来完成对远程引导设备的引导，其中将 SCSI 命令封装在 TCP/IP 协议中。iSCSI 引导典型地涉及通过一个或多个（通常是两个）iSCSI 启动器来引导 iSCSI 目标。这些 iSCSI 启动器典型地是软件 / 固件，其在计算机系统 1100 内执行并使用现有网络接口设备 1110 来经由网络 1160 访问 iSCSI 目标。iSCSI 启动器是对待自 iSCSI 目标的请求者的 / 接收者，并且 iSCSI 目标典型地在其上存储有数据，该数据对于该 iSCSI 启动器是可访问的并且是响应于接收到请求被返回的。在两个 iSCSI 启动器的情况下，启动器中的一个或二者可位于固件中或通过 PXE 来递送。同样，启动器中的一个可与计算机系统 1100 正对其进行引导的操作系统相连接。在某些实现中，并且如图 1A 内所提供，计算机系统 1100 包括 iSCSI TCP 卸载引擎 (TOE) 1115，其是代替软件 iSCSI 启动器用于与目标 /iSCSI 设备 1170 进行通信而无需网络接口设备 1110 的独立硬件设备。

[0059] 本领域的普通技术人员将意识到，图 1A 中所描绘的硬件可以进行变更。例如，作为所描绘的硬件的补充或替代，可使用诸如光盘驱动器等等之类的其它外围设备。因此，所描绘的例子意思不是暗示对于本发明的体系结构限制。图 1A 中所描绘的计算机系统可以例如是纽约阿蒙克的国际商业机器公司的产品的 IBM eServer pSeries 系统。

[0060] 在图 1A 中，当在计算机系统 1100 内没有提供 iSCSI TOE 时，软件 iSCSI 启动器被提供给设备驱动器并启动经由网络接口卡 1110 对目标 /iSCSI 设备 1170 的连接。为了使该操作，计算机系统固件加载总线驱动器。例如，操作系统中的许多当前固件安装允许在 SCSI 驱动器插入点处添加第三方 SCSI 驱动器，并且设备驱动器 1120 可以在引导序列中在该点处被加载。

[0061] 图 1B 图示示例设备驱动器体系结构和其它软件部件，其使（图 1A 中的）计算机系统 1100 能够在加载操作系统之前经由 iSCSI TOE 与网络 1160（图 1A）进行通信。在一个实施例中，设备驱动器体系结构使 iSCSI TOE 能够加载操作系统之前访问网络 1160 并连接到引导服务器设备（例如图 1A 中的 iSCSI 设备 1170）。

[0062] 在图 1B 中，计算机系统是装备有 iSCSI TOE 的 iSCSI 客户端。如所示出那样，固件 1225 提供 iSCSI 总线设备驱动器 1220，其包括直接与网络进行通信的 iSCSI (硬件) 接口设备 1215。iSCSI 驱动器 1230 包含它自己的 TCP/IP 栈用以使能通过网络对目标 /iSCSI 设备的连接以检索引导映像。iSCSI 驱动器 1230 响应于 iSCSI 引导操作生成 iSCSI 任务 1235。在系统引导期间，固件 1225 安装 iSCSI 驱动器 1230 和 iSCSI 总线设备驱动器 1220 以便能经由 iSCSI 接口设备 1215 的 iSCSI 引导操作。在该驱动器体系结构中，控制硬件接口的设备驱动器（例如，iSCSI 总线设备驱动器 1220）接收到的对于机器总线命令的响应和其它网络业务。

[0063] 通过上述示例计算机系统和网络环境，如下面所描述那样实现了本发明的各种特征。所描述的实施例引用了用于网络设备的标准支持包。在由 OpenFirmware Working Group 发表的 Open Firmware Recommended Practice: TFTP Booting Extension Version 1.0（已确认在 2006 年 8 月 22 日发表在 http://playground.sun.com/1275/practice/obptftp/tftp1_0.pdf，搜索 =% 22open % 20Firmware % 20Recommended %
20Practice% 3A% 20TFTP% 20Booting% 20Extension% 20Version% 201.0% 22) 中描述了该标准支持包。在例, 通过参考的方式入与对本发明特征的描述相关的该文档的内容。

[0064] 本发明的各实施例的具体方法是从以下理解中进行描述的, 即为了从 iSCSI 设备引导系统, 该系统的引导固定 1252 需要识别 iSCSI 硬件 (或使用 iSCSI 软件启动器), 标识 iSCSI 客户端和目标 iSCSI 设备, 与该目标建立通信, 以及从该目标加载引导文件。表示设备的固定数据结构的这种分层性质允许信息在设备与支持包之间进行传递。

[0065] 根据一个实施例, 当在计算机系统 1100 内提供 iSCSI TOE 时, 该固件将 iSCSI 启动器类型表征为类网络的设备类型, 目的是支持已扩展的网络支持包并因此使用网络服务来发现和获取目标 iSCSI 设备地址的不同分量。通过这种分层表示, iSCSI 设备 (启动器) 提供网络专用引导参数 (例如速度、模式等等) 并将那些引导参数传递给网络支持包。网络支持包获取并处理 iSCSI 设备地址中的分量并将那些变更 (也即引导参数) 传递给盘支持包。接着, 按支持包加载引导映像并将引导映像转发给 iSCSI 客户端以使能在 iSCSI 客户端处完成对操作系统引导。

[0066] 网络支持包提供对于使用在远程服务器上发现的引导文件来引导系统的支持。如在上面的参考内所进一步描述那样, 支持包还覆盖自举协议 (BOOTP) 的使用以发现引导服务器并获取使用已建立的引导变元标准的引导文件, 本发明提供对网络支持包的扩展 / 增强, 以容纳另外的新的引导变元, 该新的变元允许网络支持包通过使用发现协议来标识 iSCSI 设备地址分量。如此处所描述的实施例所提供的那样, 这些扩展 / 增强适用于 iSCSI 设备以及 “网络” 设备。

[0067] 用于网络设备的标准支持包提供对于使用在远程服务器上发现的引导文件来引导系统的支持。另外, 支持包还提供对于 BOOTP 自举协议的使用来发现引导服务器和引导文件。然而, 用于 ISCSI 设备与引导系统, 需要引导固件处理来 (a) 标识 ISCSI 客户端和目标, (b) 与目标建立通信, 以及 (c) 从目标 (或子设备) 加载引导文件。下面描述了由本发明的一个实施例提供的对网络支持包的扩展。

[0068] 由本发明的一个实施例提供的对网络支持包的一种类型的扩展使网络支持包能够容纳下述引导变元, 其允许固定用包括对于现有和未来的发现协议的使用的方法来标识 iSCSI 设备地址分量变元。这些第一扩展适用于 iSCSI 设备以及 “网络” 设备, 以适应当缺少 iSCSI TCP 减载引擎 (TOE) 硬件时由软件 iSCSI 启动器建立 iSCSI 会话的情形。

[0069] 另一种类型的扩展将一种机制添加到网络支持包以区别不同类型的自举方法 (例如, TFTP (简单文件传输协议) 自举相对于 iSCSI 自举) 以及这些方法各自的引导变元。根据本发明的一个实施例, 令在第一引导变元之前被添加的“命令关键词”来选择自举方法。多个命令关键词被定义以指定不同的自举方法和它们的相关引导变元。命令关键词定义 / 确定自举方法以及与该方法相关联的引导变元组。因为已定义的当前引导变元组对应于 TFTP 自举方法, 所以当如本发明的一个实施例所提供的那样缺少命令关键词时, 暗示着对于 TFTP 自举方法的使用。在一个实施例中, 命令关键词是互斥的。在该实施例中, 在变元中只允许一个命令关键词, 并且多个命令关键词的输入导致产生被报告的错误。接着, 各种引导变元被定义为或被需要或者可选。

[0070] 一个或多个发现协议可用于获取与所选自举方法相关联的引导变元。此处所
使用那样，发现协议定义了用于管理远程存储器建立连接的机制和可从该特定存储器获取的信息二者。例如，动态主机配置协议（DHCP）可用于获取诸如客户端地址之类的配置参数。BOOTP 协议可用于获取关于 iSCSI 目标信息，而因特网存储命名服务器（iSNS）协议和服务位置协议（SLP）可用于获取 iSCSI 客户端和 iSCSI 目标信息。根据一个实施例，使用在命令关键词之后和在第一引导变量之前插入的“限定符关键词”来选择具体的发现协议。每个限定符关键词标识公知的发现协议和通过 / 通过使用该发现协议来获取的引导变量。多个限定符关键词可在命令关键词之后被指定。如本发明所提供那样，限定符关键词的例子包括但不限于 dhcp、bootp、isns 和 slp，它们分别标识 DHCP、BOOTP、iSNS 和 SLP 协议。

提供了其中各扩展适用于网络设备的实施例，目的是适应应当缺少 iSCSI TCP 协议引擎（TOE）和使用软件 iSCSI 启动器来建立 iSCSI 会话的情形。iSCSI TOE 设备执行对 iSCSI 命令的封装并提供以太网硬件用于通过网络传送 iSCSI 命令。

因此，如前面所进一步描述那样，另外的引导变量（关键词）被建立并添加到由网络支持包用来完成以下操作的变量：(1) 区别不同类型的自举方法，也即 TFTP 自举相对于 iSCSI 自举程序，以及它们各自的引导变量，以及 (2) 指定发现协议以及 iSCSI 设备获取在 OS 引导期间所需的缺失引导参数。

iSCSI 协议将 SCSI 命令封装进网络包以便在客户端（iSCSI 启动器，例如计算机系统 1100）与 iSCSI 目标（1170）之间进行传输。标准 iSCSI 设备地址包含以下信息：

(a) iSCSI 目标服务器 (target-server) 的 IP 地址；
(b) TCP 端口号，其缺省是 3260 (target-port)；
(c) iSCSI 目标名字 (target-name) ；以及
(d) iSCSI LUN（其缺省是 0 (target-lun)）。

另外，以下附加信息可用在 iSCSI 协议中以使能在安全环境中对 iSCSI 设备的访问：

(a) CHAP 目标 ID (target-chapid)
(b) CHAP 口令 (target-chappw)。

根据本发明的实施例，当在客户端设备（例如，图 1A 中的计算机系统 1100）内提供了 iSCSI TOE 时，该 iSCSI 设备在分级固件数据结构内被表示成类网络的设备类型。由此，该 iSCSI 设备类型由与用于描述网络设备的相同类型信息（也即，机器（MAC）地址、帧大小、所支持的网络类型，等等）来描述。该类网络的设备类型表示的优点之一是，网络支持包可以得到 iSCSI 设备类型的支援。网络服务因此可用于发现并获取 iSCSI 设备地址中的不同分量。

表示设备的固件数据结构的分级性质允许信息在各设备和支持包之间传递。当存在 iSCSI TOE 设备时，该 TOE 设备提供网络专用引导参数（也即，速度、模式等等）并将这些参数传递给网络支持包。网络支持包处理这些引导变元，并依据这些变元从信息服务器获取 iSCSI 设备地址和其它参数，并且将那些参数传递给盘支持包。与 iSCSI 设备表示无关，该盘支持功能执行对引导文件的实际加载。

如前面所介绍，根据本发明的一个实施例，为了支持将 iSCSI 设备类型表示成网络类型的设备，网络支持包被扩展。不论是使用软件 iSCSI 启动器还是 iSCSI 硬件 TOE 的
描述以从目标 iSCSI 设备进行引导，对本发明的实施例的其余描述适用。

[0084] 再次参考附图，并且特别地参考图 1C，提供了根据本发明的一个示例性实施例的示例系统，该系统示图使能从 iSCSI 设备（目标）进行引导引导处理的主要部件（硬件/软件）。如所示，该系统包括逻辑地耦合到网络支持包 1320 的客户端设备（启动器，例如计算机系统 1100）的固件 1320。网络支持包 1320 接着通过逻辑地耦合到服务器 1330 和 iSCSI 设备（目标）1370。根据所描述的实施例，服务器 1330 表示可通过对应的发现协议 1325 进行访问以获取访问 iSCSI 设备 1370 所需的参数的不同服务器类型的集合。iSCSI 设备 1370 包括盘支持包 1375，其用于访问与 iSCSI 设备 1370 相关联的（或者可从 iSCSI 设备 1370 进行访问的）逻辑单元 1380。盘支持包 1375 使能基于系统固件 1300 提供给网络支持包 1320 的变量而在客户端设备 1100 处完成的远程 OS 引导操作。

[0085] 现在转到图 1D，其图示了在设备的固件初始化以及随后传送变元给网络支持包以触发远程 OS 引导操作期间，在分级结构内标识 iSCSI 设备的过程。特别地，该流程图图示了直到网络参数被传送给网络支持包之前由固件进行的设备发现过程。如方框 1402 处所示出，当系统开始引导时该过程开始。如方框 1404 所描绘，固件在搜索可能包括可选 ROM（只读存储器）的设备时遍历系统总线。这种搜索是标识系统的 1/0 拓扑的过程的一部分。在方框 1406，固件检查是否检测到可选 ROM。如果在附着到总线的任何设备中检测到可选 ROM，则该可选 ROM 被加载并且它的代码被执行，如在方框 1408 中所示出。一旦可选 ROM 代码被执行或/如果没有检测到任何 ROM，则在方框 1410 处该固件检查是否发现/检测到 iSCSI TOE 适配器。如果发现 / 检测到 iSCSI TOE 适配器，则该固件创建对该 iSCSI TOE（也即 iSCSI 设备）的表示，如在方框 1412 处所示出，并且该 iSCSI 设备表示被添加到主分级固件结构，如在方框 1414 处所示出。

[0086] 如果在方框 1410 处固件没有检测到 iSCSI TOE，则在方框 1411 处该固件检查是否检测到网络适配器。如果没有检测到任何网络适配器，则过程在方框 1415 处结束。如果检测到网络适配器，固件创建“网络”设备，如在方框 1413 处所示出，并且在方框 1414 处该网络设备被添加到主分级固件结构。所发现的 iSCSI 设备或所创建的类网络的设备使得该固件处理网络相关的引导变元（例如，速度、模式等等），如在方框 1416 中所示。紧随方框 1416，各参数被传递给在引导过程中起帮助作用的网络支持包，如在方框 1418 处所示。通过 1/0 拓扑表示的分级性质，无论 iSCSI 启动器是“网络”设备还是 iSCSI 设备，发送给网络支持包的读和写操作最终都由目标 iSCSI 设备执行。

[0087] 现在参考图 1E，其是图示网络支持包当接收到在图 1D 中的硬件初始化过程的结束处所生成和转发的引导变元时所提供的处理的流程图。特别地，图 1E 图示当选择了 iSCSI 自举方法时由网络支持包的一个实施例执行的对限定符关键词的处理。无论软件 iSCSI 启动器与“网络”设备一起被使用还是 iSCSI TOE 设备中的硬件 iSCSI 启动器被使用，与 iSCSI 自举方法相关的各参数被从设备（iSCSI TOE 或“网络”设备）传递给网络支持包。网络支持包接收到引导变元，如在方框 1502 处所示出，并且该网络支持包解析引导变元，如在方框 1504 处所提供。在方框 1506 处，网络支持包确定该引导变元是否包括指示或对应于 iSCSI 发现过程的关键词。如果该变元不包括 iSCSI 专用关键词，则网络支持包启动标准 TFTP 自举过程，如在方框 1508 处所指示。根据一个实施例，当选择了 TFTP 方法时，所提供的引导变元是：

16
[0088] [bootp, ]siaddr, filename, ciaddr, giaddr, bootp-retries, tftp-tentries, subnet-mask, blksize

[0089] 然而，根据本发明的一个实施例，需要具体的引导变元组来定义 iSCSI 命令关键词，目的是选择 iSCSI 自举方法和它对应的引导变元。因此，当选择了 iSCSI 自举方法时，提供 iSCSI 相关的参数（关键词）组，随后是盘相关变元。通过网络支持包的开放方法将这些盘相关变元传递给盘支持包。本领域的技术人员熟知盘变元，其在 Open Firmware Working Group 所发表的 PowerPC Microprocessor Common Hardware Reference Platform (CHRP) Binding 中被进一步描述。在此通过参考的方式并入该发表的相关内容。

[0090] 根据本发明的一个实施例，与 iSCSI 自举方法相关的引导变元包括以下：
iscsi, [itname = init-name], [ichapid = init-chapid, ] [ichappw = init-chappw, ]
ciaddr = initaddr, [giaddr = gateway-addr, ] [subnet-mask = net-mask, ] siaddr =
target-server, [iport] target-port, [iname = target-name, ] [ilun = target-lun, ]
[chapid = target-chapid, ] [chappw = target-chappw, ] disk-args.

[0091] 在本发明的一个实施例中，当添加限定符关键词以选择用于 iSCSI 方法的发现协议时，引导变元采取以下形式：iscsi, [dhcp = diaddr, ] [bootp, ][slp =
SLP-server, ] [isns = iSNS-server, ] [itname = init-name, ] [ichapid = init-chapid, ]
[ichappw = init-chappw, ] [ciaddr = init-addr, ] [giaddr = gateway-addr, ] [subnetmask = net-mask, ] siaddr =
target-server, [iport = target-port, ] [iname = target-name, ] [ilun = target-lun, ]
[chapid = target-chapid, ] [chappw = target-chappw, ] disk-args.

[0092] 在这些变元内，以下术语被定义 / 描述：

[0093] (a) dhcp[= diaddr] 是可选参数，其指示使用动态主机配置协议 (DHCP) 服务器来获取配置参数。diaddr 值可以被可选地提供并指定待使用的 DHCP 服务器的因特网地址。否则，执行广播查询以定位 / 找到 DHCP 服务器。取决于是否指定了限定符关键词 IPv6，用 IPv6 或 IPv4 符号表示法来表示 diaddr 值。在本说明书以下部分中更详细地描述了修改 / 增强网络支持包以支持在已扩展引导变元内使用 IPv6 和 DHCP 的方法。在此通过参考的方式并入该申请的相关内容。DHCP 服务器可提供 iSCSI 启动器地址 (ciaddr), iSCSI 目标地址 (siaddr), iSCSI 目标端口 (iport), iSCSI 目标 lun (ilun) 和 iSCSI 目标名字 (iname)。

[0094] (b) bootp 是可选的常数，其将 BOOTP 指定为“发现”协议以获取通过网络进行引导所需的参数。BOOTP 服务器可提供 iSCSI 目标地址 (siaddr), iSCSI 目标端口 (iport), iSCSI 目标 lun (ilun) 和 iSCSI 目标名字 (iname)。

[0095] (c) isns = iSNS-server 是可选的关键词值，其指示使用 iSNS 服务器来获取与 iSCSI 引导相关联的参数。iSNS-server 值是因特网存储命名服务器 (iSNS) 的因特网地址。取决于是否指定了限定符关键词 IPv6，用 IPv6 或 IPv4 符号表示法来表示该 IP 地址。iSNS 服务器可提供用于 iSCSI 引导的 iSCSI 目标名字 (iname)。给定 iSCSI 目标名字，则 iSNS 提供 iSCSI 目标服务器地址 (siaddr) 和 iSCSI 目标端口 (iport)。

[0096] (d) slp[= SLP-server] 是可选参数，其指示使用 SLP 服务器来获取引导所需的参数。SLP-server 值可被可选地提供并指定待使用的服务位置协议 (SLP) 服务器的因特网地址。否则，执行广播查询以找到 SLP 服务器。取决于是否指定了限定符关键词 IPv6，用 IPv6 或
IPv4符号表示法来表示SLP-server值。SLP服务器可提供用于iSCSI引导的iSCSI目标名字（iname）。给定iSCSI目标名字,则SLP服务器可提供iSCSI目标服务器地址（siaddr）和iSCSI目标端口（iport）。

(e) init-name = init-name 是遵从iSCSI RFC的串，其表示iSCSI启动器。iSCSI RFC1[1]指定了给定的名字在长度上可以多至233个字节。iname在iSCSI会话期间被用作事务请求者的iSCSI身份。该变元对于“网络”设备是必需参数，但是对于iSCSI设备是可选参数。如果针对iSCSI设备没有指定iname，则使用缺省的硬件值。

(f) ichapid = init-chapid 是遵从IETF的标识符（ID），其表示启动器CHAP ID并被用于帮助向目标认证启动器。特别地，启动器在iSCSI会话建立期间编码该信息并将该信息发送给目标，其中目标检查启动器实际有效。如果不需要任何认证并且不需要任何明确的启动器ID用以在发现过程中使用，则可省略ichapid。

(g) ichappw = init-chappw 是遵从IETF的口令，其表示启动器CHAP口令，并且与启动器CHAP ID一起用于帮助向目标认证启动器。如果不需要任何认证，则可省略该变元。

(h) ciaddr = init-addr 是客户端的IP地址（也即iSCSI启动器的IP地址）。取取决于是否指定了可选关键词IPv6,用IPv6或IPv4符号表示法来表示该IP地址。iSCSI启动器的IP地址在iSCSI会话期间被用作事务请求者的IP地址。它与表示iSCSI启动器的身份的iSCSI IQN串相关联。除非指定了DHCP,否则需要该参数。

(i) giaddr = gateway-addr 是iSCSI目标服务器的网关的IP地址。取取决于是否指定了可选关键词IPv6,用IPv6或IPv4符号表示法来表示该IP地址。该参数是可选的并且在 siaddr 已知并且没有在客户端的子网中的条件下可被指定。

(j) subnet-mask = net-mask 是32比特掩码，其在没有指定IPv6时定义在该特定子网上的所有IP地址的本地网络范围。

(k) siaddr = target-server 是iSCSI目标服务器的IP地址。取取决于是否指定了可选关键词IPv6,用IPv6或IPv4符号表示法来表示该IP地址。iSCSI目标IP地址在iSCSI会话期间被用作事务响应者的IP地址。如果没有指定发现协议，则需要该参数。

(l) iport-target-port 是与iSCSI目标（目标端口）相关联的TCP端口号。iport标识启动器将与哪一个目标端口进行通信。该参数可以通过指定发现协议之一来获取。该参数的缺省值是3260。

(m) iname = target-name 是遵从iSCSI RFC的串，其表示iSCSI目标。iSCSI RFC指定了给定的名字在长度上可以多至223个字节。iname在iSCSI会话期间被用作事务响应者的iSCSI身份。如果没有指定发现协议，则需要该参数。

(n) ilun = target-lun 是iSCSI目标设备的逻辑单元号（目标 lun）。该参数可以通过指定BOOTP或DHCP协议来获取。该参数的缺省值是0。

(o) chapid = target-chapid 是遵从IETF的ID，其表示目标的CHAPID并被用于向启动器认证目标。特别地，目标将在iSCSI会话建立期间编码该信息并将参数的变元发送给启动器，其中启动器将检查目标实际有效。如果不需要任何认证，则可省略该变元。

(p) chappw = target-chappw 是遵从IETF的口令，其表示CHAP口令并且与目标CHAP ID一起用于帮助向启动器认证目标。如果不需要任何认证，则可省略该变元。

根据上述的本发明的实际例，基于是 (1) 可假定缺省值还是 (2) 可使用发现协议
以获取该值，来确定对“可选的”具体标记，而对“必需的”标记适用于其值必须在引导变元组中被指定的那些参数。

[0110] 现在回到图1B，如果在方框1506处，在引导变元内存在iSCSI专用关键词，则网络支持包启动由判决方框1510，1514，1518和1522所图示的一系列检查以确定需要哪一个发现过程。尽管被图示为一个序列，但是各种确定可以由网络支持包同时而不是顺序地完成，或者作为替代这些确定可以以相对于彼此不同的次序来完成。在方框1510处，网络支持包检查变元是否指示DHCP（动态主机配置协议）发现过程。如果变元指示了DHCP发现过程，则网络支持包从已标识的DHCP服务器获取DHCP所存储的参数，如在方框1512处所示出。在由TRWF所发表的RFC2131:Dynamic Host Configuration Protocol处描述了DHCP。在此通过参考的方式并入该发表的相关内容。

[0111] 在方框1514处，网络支持包还检查变元是否指示BOOTP发现过程。如果变元指示了BOOTP发现过程，则网络支持包从已标识的BOOTP服务器获取BOOTP所存储的参数，如在方框1516处所示出。类似地，在方框1518处，网络支持包检查变元是否指示SLP发现过程，并且如果变元指示了SLP发现过程则网络支持包从所标识的SLP服务器获取SLP所存储的参数，如在方框1520处所示出。最后，在方框1522处，网络支持包检查变元是否指示iSNS发现过程。如果指示了iSNS发现过程则网络支持包从iSNS服务器获取iSNS所存储的参数，如在方框1524处所示出。

[0112] 在一个实施例中，当在变元内指定了SLP或iSNS时，则可基于“iname”在开放变元内存在或不存在而使能两个不同处理之一。因此，在开放变元中提供了iname，则所指定的iname被用于查询SLP或iSNS服务器以获取ipport和siaddr。然后，如果没有提供iname，则查询SLP或iSNS以获取可能的目标服务器的名字列表。

[0113] 通过所需额外参数的特定服务器类型和存储位置，网络支持包使用针对引导变元中所指定的服务器类型的特定发现协议来获取缺失的参数。某些iSCSI引导参数可以通过多个发现协议如DHCP和BOOTP来获取。在引导变元中所指定的所有这些发现协议可用于（同时或顺序地）获取iSCSI设备地址和其它引导参数。在一个实施例中，如果参数是自不同服务器，则将使用从最后被查询的服务器所获取的值作为替代，参数数可在获取所有引导变元时立刻停止。一旦获取了所有这些引导变元，则网络支持包使用这些额外的参数来创建引导串，如在方框1526处所示出，并且网络支持包将该引导串送传给由这些参数所标识的iSCSI设备（目标）。因此，根据本发明所描述的实施例，当需要经由特定发现过程来访问iSCSI设备时，网络支持包被扩展以使能对多个可用发现协议中的每个的iSCSI支持。

[0114] 图1F是图示在OS引导期间在iSCSI设备（目标）处发生的处理的流程图。在方框1602处，iSCSI设备接收由网络支持包创建的引导串。在方框1604处，iSCSI解释/解析引导串，并在方框1606处与客户端设备（计算机系统1100）建立iSCSI登录会话。在方框1608处，iSCSI设备检查该登录会话是否根据预先建立的操作标准进行工作。如果没有，则iSCSI生成错误，并且该引导过程在方框1610处终止。然而，假设iSCSI会话被建立并正确地工作，则该iSCSI设备将在引导变元中所找到的盘相关的参数传递给盘支持包，如在方框1612处所示出。客户端设备通过网络传送SCSI命令以从附录到目标SCSI控制器的盘获得引导文件。SCSI命令由盘支持包在方框1614处接收，而盘支持包完成从远程SCSI
盘将引导映像实际加载到客户端设备，如在方框 1616 处所示出。接着，客户端设备的 OS 引导过程完成（也即，OS 在客户端设备上执行），如在方框 1618 处所示出。

[0115] 以下例子图示由 iSCSI 自举方法内所提供的变量和特殊关键词所标识的两个不同 OS 引导过程。在第一过程中，通用网络设备在分发结构内由固件（例子 1）所标识，而在第二过程中，iSCSI 设备被标识（例子 2 和例子 3）。接着，两个过程中使用不同的变元组。由设备所处理的网络相关的参数在 iSCSI 相关的参数之前被示出。盘相关的参数在 iSCSI 相关的参数之后被示出。

[0116] 例子 1：通过本例，BOOTP 服务器用于获取 siaddr、iport 和 iname。未被指定的其它变元取缺省值。同样，ciaddr 被用 IPv6 格式表示。在本说明书下部分中更详细地描述了用于通过 iSCSI 协议使用 IPv6 的过程。

[0117] open network-device[promiscuous, speed = 100, duplex = full, iscsi, ipv6, bootp, itname = init-name, ciaddr = init-addr, 2
[0118] 例子 2；通过当前例子，SLP 服务器提供 siaddr 和 iport 用于指定的 iname。再次，未被指定的其它变元取缺省值，并且 ciaddr 被用 IPv6 格式表示。

[0119] open iscsi-device[speed = 100, duplex = full, iscsi, ipv6, sip, iname = target-name, ciaddr = init-addr, 2
[0120] 例子 3；通过该第三例子，diaddr DHCP 服务器提供 siaddr、ciaddr、ilun、iport 和 iname，而 diaddr 被用 IPv6 格式来表示。

[0121] open iscsi-device[duplex = full, iscsi, ipv6, dhcp = diaddr, chapid = target-chapid, chappw = target-chappw, 2
[0122] 本发明提供了一种方法和系统，用于使得 InfiniBand（IB）主计算机设备能够通过 IB 网络从远程引导设备引导操作系统。增强设备的固件，以便在表示计算机设备的硬件的层次数据结构内提供 IB 主机通道适配器（HCA）的表示。提供了 “ib- 引导” 支持包来协助封装用于使用 ib 端口设备通过 IB 网络进行引导的变元，并且定义引导变元集来由所述 ib- 引导支持包利用。ib- 引导支持包对指定目标 IB 端节点以及一个或多个关键字的引导变元进行处理。

[0123] 利用固件结构的层次特性，基于指定关键字来将这些 IB 专用变元传递至适当的支持包，以便完成引导处理。在一个实施例中，提供了两个关键字，各自标识网络支持包和套接字直接协议（SDP）支持包中特定的一个。引导平台使用 IB 结构来连接至 LAN NIC 以获得对该 LAN 中的引导服务器的访问，或者连接至 IB 结构中的端节点。第一关键字指定将要实现 IPolB（基于 IB 的网际协议（IP））网络引导方法，由此使用 IPolB 网络协议来到达引导服务器并且找到引导文件。第二关键字指定将要实现 IB 网络引导方法，由此使用 SDP 网络协议来连接至 IB 结构中的引导服务器并且找到引导文件。因而通过支持包之一提供了对引导服务器的访问，并且经由 IB 网络将引导映像返回给主机设备用于完成引导操作。

[0124] 现参照附图，图 2A 根据本发明描述的实施例，提供了利用 InfiniBand（IB）网络连接来与各种外围和组网设备（包括一个或多个远程引导服务器）进行通信的组网计算机系统的主要硬件组件的高级表示。如所示出的，计算机系统 2100（文中可互换地称为 IB 网络内的主计算机或主计算机设备）包括经由存储总线 2103 搭配于存储器 2102 的中央处理器 (CPU) 2101，存储总线 2103 又通过输入 / 输出 (I/O) 总线接口 2105 提供到各种其它外围或
网络设备的连接。CPU 2101 是执行来自存储器 2102 的指令以及处理来自存储器 2102 数据的通用可编程处理器。主存储器 2102 可以是动态随机访问存储器 (DRAM), 在其中，从存储器 (2123/2124) 加载数据用于由 CPU 2101 进行处理。

[0125] 存储总线 2103 提供了用于在 CPU 2101、主存储器 2102 和 I/O 总线接口 2105 之间传送数据的数据通信路径。I/O 总线接口 2105 进一步耦合于系统 I/O 总线 2104, 其用于传送数据到各种 I/O 单元和传送来自各种 I/O 单元的数据。I/O 总线接口 2105 通过系统 I/O 总线 2104 与多个 I/O 接口单元 2111-2113 进行通信。I/O 接口单元 2111-2113 也被称为 I/O 处理器 (IOP) 或 I/O 适配器 (IOA)。系统 I/O 总线 2104 可以是工业标准 PCI 总线或任何其他适当的总线技术。I/O 接口单元支持与各种存储器和 I/O 设备进行通信。举例来说，终端接口单元 2111 支持连接一个或多个用户终端 2121-2122。存储接口单元 2112 支持连接一个或多个直接访问存储设备 (DASD) 2123-2124 (其可以表示独立磁盘冗余阵列 (RAID) 存储器，被配置作为主计算机的单个大型存储设备)。其它类型的 I/O 设备（未示出）可以依附于 I/O 接口单元，例如磁带驱动、打印机和网络收发机。

[0126] 根据本发明的一个实施例，I/O 接口单元之一是 IB I/O 适配器 /HCA2113，由于 IB I/O 适配器 /HCA 2113 使得主计算机 2100 能够连接至 IB 网络，因此文中可互换地将 IB I/O 适配器 /HCA 2113 称为主机通道适配器 (HCA)。HCA 是可编程 DMA（直接存储器访问）引擎，其生成和消耗分组。

[0127] IB I/O 适配器 2113 支持到 InfiniBand 网络 2130 的连接，InfiniBand 网络 2130 包括交换机 2131 和多个点到点连接 2132-2135。InfiniBand 网络可以用于连接至各种设备的本地网。如所知道的，InfiniBand 网络 2130 提供到其它计算机系统（例如，经由目标通道适配器 (TCA) 2136 耦合于交换机 2131 的 IB-引导服务器 2137）的连接。InfiniBand 网络 2130 还提供到路由器或适配器的连接，而路由器或适配器又将 InfiniBand 网络 2130 耦合到其它局域或广域网络。在图 2A 的示例性实施例中，TCA 2139 经由多网适配器 2140 连接至局域网 (LAN) 2145。LAN 2145 提供对远程引导服务器 2150 的访问。此外，路由器 2138 连接至因特网 2146，其提供对第二远程引导服务器 2151 的访问。如文中所提供的，除了 InfiniBand I/O 适配器 (HCA) 2113 和相连的 IB 网络 2130 以外，文中将计算机系统 2100 的组件也称为“主计算机”，以便区别需要来自位于跨 IB 网络 2130 的引导服务器的引导 OS 的主机设备。如所说明的，提供了两个外部网络，即 LAN 2145 和因特网 2146。将通过主机设备 2100 对因特网 2146 和 LAN 2145 的访问描述为发生在 IB 网络 2130 的“端节点”，例如 TCA 2139。

[0128] 尽管图 2A 中所示的 IB 网络 2130 具有单个交换机 2131 和对于相应的数字设备的四个连接 2132-2135，然而应当理解，IB 网络可以具有不同数目的交换机、适配器、路由器和链路（但可能不具有交换机或路由器）。IB 网络的拓扑可以是显著不同并且更加复杂的，并且因而不同数目的类型设备可以耦合于 IB 网络。

[0129] 应当理解，图 2A 旨在高级地描绘计算机系统 2100 的代表性主要组件，并且各个组件可以具有与图 2A 中所表示的组件更大的复杂性。应当进一步认识到，可以给出不同于或者除了图 2A 中所示的那些组件以外的组件，并且这样的组件的数目、类型和配置是可以改变的。因而，仅仅通过举例的方式给出图 2A，并并不旨在隐含对本发明的任何限制。

[0130] 本发明通过跨 IB 网络 2130 发生的引导过程，为连接至 IB 网络 2130 的计算机设
备 2100 启用了远程引导过程。在一个实施例中，在 IB 结构上支持完成网络引导过程的两
种不同方法，即 (a) LAN 类型网络引导和 (b) IB 网络引导。LAN 类型网络引导方法涉及连接
至位于依附于远程 IB 端节点的 LAN 中的引导服务器，以及使用在 IB 结构上发送的传输控
制协议 / 网际协议 (TCP/IP) 命令（在此称为 IPoIB）。IB 网络引导方法涉及连接至位于远
程 IB 端节点处 / 中的引导服务器，以及使用诸如套接字直接协议 (SDP) 的传输协议来获得
引导文件。

[0131] 另外，如在本说明书中描述的，在远程引导操作的全面实施中，可以通过计算机设
备 2100 来支持存储器引导方法。存储器引导方法涉及连接至直接或通过存储区域网 (SAN)
而依附于远程 IB 节点的存储设备，以及通过 IB 结构发送存储协议命令，例如 SCSI（小型计
算机系统接口）RDMA（远程直接存储器访问）协议 (SRP)。在本说明书中提供了对该实施方
式的详细说明。

[0132] 在以上示例计算机系统和网络环境中，本发明的第一方面涉及修改系统固件，以
便使得该固件能够标识和表示在系统初始化期间系统固件按照使得 IB 端节点用于引导操
作的方式所生成的层次结构内的 IB 端节点。在计算机系统通电（或重新引导）之后，系统
经历引导过程，通过该引导过程，系统固件通过标识和建立 / 启动系统中的设备来准备用于
操作的系统。在引导过程期间，固件将系统硬件表示为具有附属设备的互联总线的层次。该
层次描述了用户配置选择，并且含有用于硬件设备的固件设备驱动器以及由那些驱动器所支
持的路由器。

[0133] 在 IB 网络内，在本发明的一个实施例中，增强主机系统的引导固件以便识别存在
IB HCA 以及一个或多个 IB 端节点设备。根据一个实施例，新的设备类型定义为“hca”设
备类型，其表示 IB HCA。hca 设备的特征在层次固件数据结构中通过节点全局唯一标识符
来描述。hca 设备含有开启和关闭设备的接口以及从设备读取和写入设备的接口。hca 设
备包括若干“子设备”，其被称为“ib 端口”设备并且由端口全局唯一标识符来标识。除了
为父 hca 设备定义的开启、关闭、读和写接口以外，ib 端口设备还包括使该设备能够作为引
导设备来使用的加载接口。

[0134] 本发明的第二方面涉及命令 / 限定符关键字的创建和利用，该命令 / 限定符关
键字被添加到在系统初始化期间由系统固件所生成的引导变元。命令关键字指定分组传
送协议和关联的变元集。通过 ib- 引导支持包来理解这些命令关键字，ib- 引导支持包接收
来自系统固件的引导变元。在本发明的一个实施例中，通过在用于 ib 引导支持包的开启接
口的变元中使用命令关键字来选择特定的网络引导方法。命令关键字通过指定实际的具
体网络引导方法来修改与自举方法关联的引导变元。根据一个实施例，命令关键字是引导
变元中的一变元，并且每次只可以指定一个命令关键字。如以下进一步描述的，这些命令
关键字是互斥的，并且因而，仅将限定符关键字之一添加到标准引导变元之前来启用 IB 级
远程引导操作。如果指定了超过一个的命令关键字，则发送具有警告消息的失败结果给主
计算机设备的用户，并且该设备不开启。

[0135] 因而，在本发明的说明性实施例中，创建了以下两种命令关键字，且提供其相应的
定义为：

[0136] (a) ipoib：其指定 IPoIB 网络引导方法。当指定 ipoib 时，使用 IPoIB 网络协议来
访问引导服务器和找到引导文件。引导平台使用 IB 结构来连接至 LAN NIC（网络接口卡）
以获得对 LAN 中的引导服务器的访问;以及

(b) sdp; 其指定 IB 网络引导方法。当指定 sdp 时, 使用 SDP 网络协议来访问 IB 结
构中的引导服务器。

在以上引导变量内, IB 专用变量及其描述包括:

(a) iocguid = IOCGUID: 其是标识目标 I/O 控制器或 I/O 单元的 64 位惟一
EUI-64 顺博标识符 (compliant identifier); 以及

(b) portgid = IOCGID: 其是标识目标 I/O 单元中的端口的 128 位数。以 IPv6 格
式表示该变量并且该变量由 64 位 GID 前缀和分派给 I/O 控制器的 64 位端口 GUID 组成。

如文中所提供的, IOCGUID 和 IOCGID 唯一标识 InfiniBand 结构中的端节点。

ib-引导过程建立与目标 I/O 控制器或端节点的连接, 并且然后利用通过命令关键字所指
定的协议在 IB 传输上发送分组。在本说明书中还描述了网络相关的变量 (network-args)
的例子。

如在下面详细描述的, 本发明的一个实施例与增强的网络支持包进行操作, 该网
络支持包支持对限定符关键字的引导变量的附加, 例如 ipv6 和 dhcp。视情况在第一引导变
元之前添加 ipv6 关键字以指示使用 IPv6 格式来表示引导变量内所提供的任何 / 所有网络
地址。由于 IB 网络以 IPv6 格式传输地址, 因此当随引导变量提供命令关键字 IPoIB 时,可
以假定 ipv6 关键字为缺省地址格式。

本发明的第三方面涉及利用 ib- 引导支持包来使得能够无缝连接到来自远程引
导设备的 SDP 网络引导过程或 IPoIB 网络引导过程。ib- 引导支持包实现通常由 hca 设备
驱动器利用的特定分组传送协议, 当在 IB 上引导时利用该特定分组传送协议。

ib- 引导支持包协助封装用于使用 ib 端口设备在 IB 网络上进行引导的变量, 并
且定义引导变量集来由 ib- 引导支持包利用。ib- 引导支持包处理指定了目标 IB 端节点
(即, 褚合于 IB 结构的 TCA 或路由器或适配器, 以及直接或经由网络连接到 IB 结构的引
导服务器) 的引导变量。在一个实施例中, ib- 引导支持包还实现通常由 hca 设备驱动器利
用的特定分组传送协议, 当在 IB 上引导时利用该特定分组传送协议。

利用固件结构的层次特性, 基于如 ib- 引导支持包所处理的指定关键字, 将这些
IB 专用变量传递至适当的支持包, 以便启动对引导映像的检索。在一个实施例中, 提供了两
个关键字, 各自标识网络支持包和套接字直接协议 (SDP) 支持包中特定的一个。引导平台
使用 IB 结构来连接至 LAN NIC 以获得对 LAN 中的引导服务器的访问, 或者连接至 IB 结构
中的端节点。第一关键字指定将要实现 IPoIB 网络引导方法, 由此使用 IPoIB 网络协议来
到达引导服务器并且找到引导文件。第二关键字指定将要实现 IB 网络引导方法, 由此使用
SDP 网络协议来连接至使用 SDP 的 IB 结构中的引导服务器。

因而,取决于随引导变量提供的关键字, ib- 引导支持包可以将引导变量传递至网
络支持包来完成网络引导方法。可选地, ib- 引导支持包可以将引导变量传递至 SDP 支持
包来完成 IB- 网络引导。根据一个实施例, 用于 IPoIB 引导的引导变量如下:

open ibport-device:iocguid = IOCGUID, portgid = IOCGID, ipoib, network-args

当引导变量指定 IPoIB 网络引导方法时, ib- 引导支持包处理 IB 专用引导变量并
且将网络相关的变量传送至网络支持包。

23
[0149] 图 2B 是根据本发明的一个实施例描绘了一过程的流程图，通过该过程，系统组件发现 IB HCA 并且表示层次拓扑内的设备及其子设备。参照图 2A 所示的系统组件描述了图 2B 的过程。过程从块 2202 开始，在此，在诸如主机系统 2110 (图 2A) 的主机系统中，由诸如系统组件 2145 (图 2A) 的组件启动硬件初始化。如块 2204 所示，在系统初始化期间，系统组件 2145 通例流系统组件并且标识出主机系统 2100 的 I/O 拓扑。在判定块 2206 处，如果在附于子线的任何设备中检测到选项 ROM，则加载该选项 ROM 并且执行其代码，如块 2208 所示。作为执行代码的结果，创建设备的表示并且将其添加到主层次 1/O 拓扑，如块 2214 所示。

[0150] 在判定块 2206 处，如果未检测到选项 ROM，则在块 2210 处系统组件 2145 确定是否检测到 IB HCA (或“hca”设备)。如果检测到 IB HCA 设备，则系统组件 2145 创建 hca 设备及其子 IB 端口设备的表示，如块 2214 所示。系统组件 2145 还创建 1/O 拓扑内 IB 端口设备的位置的设备表示，如块 2214 处所指示的。因为“ib 端口”设备是利用特定的引导变量集开启的，所以系统组件 2145 获得开启 ib 端口设备所需要的引导变量。系统组件然后将引导变量传递至 ib- 引导支持包用于处理，如块 2218 所指示的。

[0151] 现转至图 2C，其根据本发明的一个说明性实施例，提供了示例系统的框图表示，该示例系统具有这样的主要组件（硬件和软件应用体），其使得能够通过 IB- 引导支持包来利用 IB- 引导变量，从而启动来自远程引导设备的 IBoIP 引导或 SDP 引导。如所示出的，正被引导的设备的系统组件 2145（例如，主机系统 2100（图 2A））在通信上耦合于 ib- 引导支持包 2310。ib- 引导支持包建立与引导变量中所指定的目标 1/O 控制器的连接。当从 IB 连接的引导设备引导时，ib- 引导支持包 2310 传输 SDP 引导变量至 SDP 支持包 2330。其利用 SDP 网络协议获得来自引导服务器 2137（图 2A）的引导文件。当通过 IB 网络连接的服务器的引导时，ib- 引导支持包 2310 将 IPoIB- 引导变量转发至增强的网络支持包 2320。

[0152] 网络支持包 2320 又经由 IP 网络在通信上耦合于引导服务器 2150（图 2A）。根据一个实施例，网络支持包 2320 进一步连接至多个其它的服务器，例如 DHCP 服务器和 BOOTP 服务器（未具体示出），这些服务器表示了可以通过对应的发现协议来访问以便获得用于访问引导服务器 2150 所需要的附加参数 / 变元的不同服务器类型的集合。

[0153] IP 引导服务器 2150 和 IB 引导服务器 2137（图 2A）能够将引导映像提供给计算机系统 2100 的系统组件 2145，以便使得计算机系统 2100 能够完成引导操作。服务器 2150 和 2137 分别从网络支持包 2320 和 SDP 支持包 2330 接收相应的引导变量集。然而，网络支持包 2320 通过 IPoIB 消息接收变量并且传输 IPoIB 消息，这与 SDP 支持包 2330 所支持的消息传输的类型 / 格式（即 IB 信息）不同。

[0154] 根据说明性实施例，为增强的网络支持包 2320 提供了一种机制，该机制使得网络支持包 2320 能够接收包括关键字（包括 ipv6 和 dchp 在内）的引导变量，这是因为 IB 通过 IPv6 寻址方式进行操作。更适合而言，增强的网络支持包 2320 提供对 IPv6 地址格式的支持并且使得能够经由 IPoIB 消息检索引导映像。在本说明书中较为详细地描述了使得网络支持包 2320 能够利用 IPv6 地址格式的方法。

[0155] 图 2D 描绘了在 ib- 引导支持包与根据包括在引导变量内的命令关键字所选择的其它两种支持包之间的引导变量的流程。ib- 引导支持包提供指定了目标 IB 端节点的引导变量，利用组件结构的层次特性将这些 IB 专用变量传递至适当的支持包以便完成引导处理。过程开始于块 2402，其说明了 ib- 引导支持包从系统组件接收引导变量。ib- 引导支持
说明书

如果引导变元未指示 IPoIB 引导, 那么在块 2412 处进行引导变元是否指示 SDP 引导的下一确定。如果指示 SDP 引导, 则 ib- 引导支持包生成 SDP 引导所必要的引导串, 从块 2414 所示, 并传递生成的引导串转发至 SDP 引导支持包, 块 2416 所示的。因而, 当选择 IB 网络方法时, ib- 引导支持包向 SDP 支持包或 InfiniBand 键路上促进或提供直接文件传输协议的任何其它的包传送控制。否则, 如果引导变元既未指示 IPoIB 引导, 也未指示 SDP 引导, 则报告错误, 如块 2418 处所示, 并且引导过程结束。

图 2E 是根据本发明的一个实施例描绘的对引导变元的处理的流程图, 当利用在 IB 键路 (IPoIB) 上传输 TCP/IP 消息来加载来自 LAN 服务器 (例如图 2A 的引导服务器 2150) 的引导文件时, 网络支持包 2320 进行对引导变元的这一处理。通过 I/O 拓扑表示的层次特性, 由网络支持包所进行的读和写操作被传递至 ib- 引导支持包 2310 并且实际上是通过 HCA 设备 (例如图 2A 的 HCA 设备 2113) 实现的。HCA 2113 封装 TCP/IP 命令并且经由 IB 键路 (例如图 2A 的 IB 总线 2133) 将命令发送至指向的 IB 目标节点 (例如图 2A 的 TCA 2139)。如果 TCP/IP 命令的目标不是目标 IB 节点, 则目标 IB 节点中的 LAN 驱动器转发分组到 LAN 2145 (图 2A) 内适当的目的地 (例如, 引导服务器 2150)。

该过程开始于块 2502, 其说明了诸如图 2C 的网络支持包 2320 这样的增强网络支持包 (从图 2C 的 ib- 引导支持包 2310) 接收引导变元。增强的网络支持包 2320 解析引导变元, 从块 2504 所示, 并且在块 2506 处确定引导变元是否含关键字 ipv6。当关键字 ipv6 含于引导变元内时, 网络支持包 2320 配置所有的寻址功能来支持 IPv6 格式的地址, 如块 2508 所提供的。

然后在块 2510 处, 网络支持包 2320 确定引导变元是否含有关键字 dhcp。如果关键字 dhcp 含于引导变元内, 则网络支持包 2320 发送并接收 IPoIB 消息, 以便从 DHCP 服务器获得引导信息 (即寻址和其它参数), 如块 2512 处所示。关键字 dhcp 可以具有变元 “diaddr”, 其将增强的网络支持包 2320 定向到准确的 DHCP 服务器, 将从该 DHCP 服务器获得引导服务器地址和其它参数。可选地, 当没有随关键字 dhcp 提供变元时, 网络支持包 2320 发布广播以便在网络上找到 DHCP 服务器。

当在块 2514 处确定完成了从 DHCP 服务器检索引导参数之后, (或者当 dhcp 没有包括在引导变元内时), 网络支持包 2320 在块 2516 处确定在引导变元内是否指示了自举过程 (bootp)。当 bootp 是引导变元之一时 (即在以上关键字之一或二者之后, 或者没有任一个关键字出现在引导变元内), 网络支持包 2320 从 BOOTP 服务器检索寻址和其它参数作为利用 IPoIB 消息的自举操作, 如块 2518 所说明的。一旦自举操作完成, 如在判定块 2520 处所确定的 (或者如果在引导变元内没有包括自举变元), 则网络支持包 2320 生成 / 创建引导串, 如块 2522 处所示。然后网络支持包 2320 在 IB 上利用 TFTP 消息来从远程服务器加载引导映像, 如块 2524 所示。然后固件在计算机系统 2100 (图 2A) 处完成引导过程, 如块 2526 所指示的。
图 2F 是描述了当利用 SDP 协议来从 IB 端点（例如图 2A 的服务器 2137）加载引导文件时，由诸如图 2C 的 SDP 支持包 2330 这样的 SDP 支持包所进行的对引导变元的处理的流程图。通过 1/O 拓扑表示的层次特性，由 SDP 支持包所进行的读和写操作被传递至 IB 引导支持包并且实际上是由 HCA 设备实现的。HCA 设备封装 SDP 消息并且经由 IB 链路将消息发送至指定的 IB 目标节点。

该过程开始于块 2602，其说明了 SDP 支持包 2330 从 ib- 引导支持包 2310 接收 ib- 引导变元。SDP 支持包 2330 解析该引导变元，如块 2604 所示，并且然后通过 IB 将 SDP 命令发送至引导服务器（例如图 2A 的引导服务器 2137），如块 2606 所示。在块 2608 处进行是否已经完成 SDP 命令的传输的确定，并且当该传输已经完成时，引导服务器 2137 提供引导映像，其被加载到主计算机 2100（图 2A）上，如块 2610 所示。然后，在主计算机 2100 处着手并完成引导操作，如块 2612 所指示的。

该发明提供了一种扩展计算机网络系统的网络支持包以接收 IPv6 格式的网络地址和允许使用诸如 DHCP 之类的其他引导发现协议以便获取使用远程服务器上存储的引导文件引导系统所需的信息的方法和系统。

现在参考附图，图 3A 是示出了通过网络 3160 连接到远程 iSCSI 设备（3170）（例如，引导服务器或目标）以及一个或多个信息源 3180（例如，DHCP 服务器或 BOOT 服务器）的计算机系统 3100 的结构表示的方块图。贯穿此说明，将 iSCSI 设备 3170 可互换地称为“目标”或“iSCSI 服务器”或“引导服务器”。计算机系统 3100 还可以称为“启动器”或“客户端”或“客户端设备”。计算机系统 3100 可以是个人计算机、服务器、移动或手持电子设备或其他数据处理机器。网络 3160 可以是局域网、域域网、广域网或其他类型的机器网络。网络 3160 可以是有线或无线网络，并且可以使用诸如网际协议（IP）之类的无连接分组交换协议和 / 或其他网络协议。

计算机系统 3100 包括将计算机系统 3100 连接到网络 3160 的网络接口设备 3110。网络设备 3110 表示计算机系统 3100 的硬件组件，并且可以是网络接口卡（NIC）或集成网络设备（例如，内置在诸如计算机系统 3100 的母板之类的主板板内的网络适配器）。

设备驱动器 3120 控制网络接口设备 3110 以提供通道。操作系统（OS）3125 以及计算机系统 3100 的应用层中的一个或多个软件应用 3150 可以通过该通道与网络 3160 通信。设备驱动器 3120 允许使用网络接口设备 3110 通过网络 3160 执行数据存储操作 3130。例如，设备驱动器 3120 可以包括允许 SCSI 命令通过网络 3160 与存储目标（iSCSI 设备 3170）交互的 iSCSI 启动器。

至少一个设备驱动器 3120 可以支持通过网络 3160 从 iSCSI 设备 3170 引导，在此情况下 iSCSI 设备 3170 表示远程引导设备。作为远程引导设备，iSCSI 设备 3170 在借助通过网络 3160 完成的一系列过程将（操作系统的）引导映像加载至计算机系统 3100 之前进行初始化。iSCSI 设备 3170 可以位于靠近或远离计算机系统 3100 的物理空间。例如，iSCSI 设备 3170 可以通过局域网 3160（例如，吉比特以太网）连接到计算机系统 3100，并且可以与计算机系统 3100 位于同一建筑物，机架或机箱中。

此外，设备驱动器 3120 允许使用网络接口设备 3110 通过网络 3160 执行网络业务操作 3140。至少一个设备驱动器 3120 可以将自身作为网络设备驱动器呈现给操作系统并支持通用网络业务。因此，操作系统可以查看其中一个设备驱动器 3120 并与其交互，好像此驱
动器是在计算机系统的引导过程中的总线驱动器和存储驱动器之后加载的常规网络设备驱动器。支持的网络业务可以包括多个不同的组网协议，例如 TCP/IP、ODI（开放式数据链路接口）、PXE（预引导执行环境）以及各种其他标准化和 / 或基于公司 / 操作系统的网络协议，具体取决于实施方式。通常，设备驱动器 3120 允许计算机系统 3100 使用多个组网协议与一个或多个信息源 3180 通信。

[0169] 虽然计算机系统 3100 的设备驱动器 3120 在图 3A 中出现在与操作系统和硬件分离的层中，但是应理解，所述设备驱动器可以在计算机系统 3100 内的不同上下文中运行，具体取决于实施方式和操作状态（从关闭电源到引导序列，直至到达运行模式，此时操作系统完全加载并且控制计算机系统 3100）。此外，在此所述的各种驱动器功能可以在硬件、固件和 / 或软件中实现。

[0170] 如上所述，设备驱动器 3120 可以支持引导到远程引导服务器（目标 iSCSI 设备 3170）并支持通用网络业务，而至少一个设备驱动器 3120 将自身作为网络设备驱动器呈现给固件。可以使用 iSCSI 协议完成引导到远程引导设备，其中将 SCSI 命令封装在 TCP/IP 协议中，并使用为固件层次结构内的网络型设备的软件 iSCSI 启动器或 iSCSI TOE 适配器将其发送到目标设备。因此，虽然从标准网络设备过程的角度描述和说明了本发明，但是所述的本发明的特性同样适用于 iSCSI 引导过程，其中远程引导服务器（目标 iSCSI 设备 3170）是 iSCSI 设备。在本申请的相关方面提供了对 iSCSI 引导过程的具体描述。

[0171] 本领域的技术人员将理解，图 3A 中示出的硬件可以有所变化。例如，除所述硬件以外或替代所述硬件，还可以使用诸如光盘驱动器之类的其他外围设备。因此示出的例子并非旨在暗示对本发明的体系结构限制。图 3A 中示出的计算机系统可以例如是 IBM eServer pSeries 系统，其是位于纽约阿蒙克的国际商业机器公司的产品。

[0172] 在上述示例计算机系统和网络环境中，向网络支持包添加了某种机制以允许支持其他引导发现协议和其他网络地址格式。所述机制定义了添加到由系统固件生成的引导变元并由网络支持包处理的限定符关键字。在一个实施例中，所述限定符关键字通过指定获取其他参数 / 变元（即，通过特定的引导发现协议）、参数 / 变元格式以及与远程引导过程关联的其他参数的方法，来修改与自举方法关联的引导变元。可选地在与所述自举方法关联的变元之前添加限定符关键字以便修改引导变元。一次可以向引导变元添加多个限定符关键字，并且可以按照关键字在所述变元中的顺序来应用所述关键字。

[0173] 以下提供了在本发明之前与标准自举操作一起使用的当前 / 现有引导参数。然后为每个变元提供了其各自的定义和 / 或说明。因此，现有的引导变元提供了:

[0174] [bootp,]siaddr, filename, ciaddr, giaddr, bootp-retries, tftp = retries, subnet-mask, blksize

[0175] 在此引导变元中，所示术语的定义和用法包括:

[0176] siaddr，其是 TFTP 服务器的 IP 地址；

[0177] filename，其是要由 TP 从服务器加载的文件的文件名。

[0178] ciaddr，其是客户端（即，被引导的系统）的 IP 地址；

[0179] giaddr，其是 BOOTP “网关” 的 IP 地址，并且它不一定与网络网关相同；

[0180] bootp-retries，其是在确定 BOOTP 过程已失败之前尝试的最大重试次数；

[0181] tftp-retries，其是在 TFTP 过程停止之前尝试的最大重试次数；

27
**subnet-mask**，其是定义此特定子网上所有 IP 地址的本地网络范围的 32 位掩码；以及
**blksize**，其是指定 TFTP 读取或写入请求的块中的八位组数的十进制数。

上述为网络支持包提供的变元指定了用于通过网络引导计算机系统的 TFTP 自举方法。TFTP 自举方法的引导变元提供了指定 BOOTP 协议以查找引导服务器和引导文件的选项。如果未指定 BOOTP 协议，则可以使用具体包括 RARP/APR 的其他协议来获取引导文件。

本发明的一个实施例扩展了上述引导变元以包括其他关键字和关联的变元（如果适用）。在示例性实施例中，由修改后的网络支持包支持的示例扩展引导变元包括以下内容:

**[ipv6, [dhep[ = diaddr, ]]**
**[bootp, ]siaddr, filename, ciaddr, giaddr, bootp-retries, tftp-retries, subnetmask, blksize.**

如扩展引导变元的上述示例性实施例中所提供的，在扩展的引导变元中定义了两个限定符关键字（以粗体表示），即 **ipv6** 和 dhcp，并且 dhcp 具有相应的变元 diaddr。可选地将 **ipv6** 关键字添加在第一个引导变元之前以指定使用 IPv6 格式来表达引导变元中提供的任何 / 所有网络地址。在示出的实施例中，没有关键字 “ipv6” 指示以 IPv4 格式表达所有网络地址。因此，IPv4 仍为默认的网络地址格式并且不必专门添加在变元之前。根据所述的实施例，关键字 **ipv6** 位于变元中使用的任何其他限定符关键字之前。

**dhcp** 关键字用于指示将使用动态主机配置协议 (DHCP) 服务器以获取引导参数。可以可选地提供 diaddr 值以指定要使用的 DHCP 服务器的因特网地址。否则，如果未提供 diaddr 值，则网络支持包将发出广播查询以便在网络上查找 DHCP 服务器。diaddr 值以 IPv6 或 IPv4 表示法表达，具体取决于是否指定了可选的限定符关键字 **ipv6**。当被访问时，DHCP 服务器可以提供用于 TFTP 自举的客户端地址（ciaddr）、服务器地址（siaddr）以及文件名。在由 IETF 发布的 RFC 2131；动态主机配置协议中提供了如何从 DHCP 服务器获取这些和其他变元的说明。

图 3B 是根据本发明的一个示例性实施例的示例系统的方块图表示，所述系统具有主要组件（硬件和软件实用程序），其通过修改 / 增强后的网络支持包 3220 利用扩展引导变元以便使用 IPv6 地址格式和备选协议来查找诸如 DHCP 的引导参数。如所示出的，所述系统包括例如计算机系统 3100 的客户端设备（其为 iSCSI 环境中的启动器）的固件 3200，所述固件可通信地连接到修改 / 增强后的网络支持包 3220。网络支持包 3220 又可通过通信地连接到服务器 3230，后者包括 DHCP 服务器 3230A、BOOTP 服务器 3230B 以及引导服务器（iSCSI 设备）3270（其为 iSCSI 环境中的目标）。根据所述的实施例，服务器 3230 可以由相应引导发现协议 3225 访问以获取访问引导服务器 3270 所需的参数的不同服务器类型的集合。

当在 iSCSI 环境中实现时，引导服务器 3270 包括用于访问与引导服务器 3270 关联（或可从引导服务器 3270 访问）的逻辑单元 (LUN) 的盘支持包。对于此备选实施方式，盘支持包允许根据系统固件 3200 提供给网络支持包 3220 的引导变元和 / 或网络支持包 3220 从服务器 3230 检索的参数，来检索引导映像以供客户端设备 3100 使用。

根据本发明的一个实施例，为修改 / 增强后的网络支持包 3220 提供了一种机制,
其使得网络支持包能够接收包括关键字（包括 ipv6 和 dhcp）的引导变元。更具体地说，修改 / 增强后的网络支持包 3220 提供对 IPv6 地址格式的支撑，并且能够从 DHCP 服务器检索引导服务器地址和其他信息。

[0193] ipv6 和 dhcp 扩展同样适用于网络引导和 iSCSI 引导，后者是涉及访问目标服务器中的盘（LUN）以获取 OS 映像的另一种自举方法。但是，如下面进一步描述的，对于网络引导，TFTP 是用于完成网络引导的自举方法，并且当未向引导变元增加 iSCSI 关键字时，TFTP 是默认的自举方法。在本申请的相关部分中描述了向引导变元增加 iSCSI 关键字。可以使用 TFTP 自举方法或 iSCSI 自举方法，并且 ipv6 和 dhcp 的扩展适用于网络引导过程以及 iSCSI 引导过程。

[0194] 图 3C 是示出了根据本发明的一个实施例的网络支持包借助其使用扩展引导变元以便使用 IPv6 寻址格式以及指定备选协议以便查找引导参数（如 DHCP）的过程的流程图。过程开始于方块 3302，方块 3302 表示出了使用 / 增强后的网络支持包 3220（从系统固件）接收引导变元。增强后的网络支持包 3220 解析引导变元（如方块 3304 所示）并在方块 3306 判定引导变元是否包含关键字 ipv6。当引导变元中包含关键字 ipv6 时，网络支持包 3220 将处理 IPv6 格式的地址并使用 IPv6 格式的地址执行寻址功能，如方块 3308 所示。

[0195] 然后在方块 3310, 网络支持包 3220 判定引导变元是否包含关键字 dhcp。如果引导变元中包含关键字 dhcp，则网络支持包 3220 将从 DHCP 服务器获取寻址和其他参数，如方块 3312 所示。如上所述，关键字 dhcp 可以包括变元“diaddr”，此变元将网络支持包 3220 导向从其获取引导服务器地址和其他参数的准确的 DHCP 服务器。备选地，当变元不具备关键字 dhcp 时，网络支持包 3220 将发出广播以在网络上查找 DHCP 服务器。

[0196] 在从 DHCP 服务器检索各种寻址参数之后（或者当引导变元中不包含 dhcp 时），网络支持包 3220 将在方块 3314 判定引导变元中是否指示了其他引导发现协议。当 BOOTP 是引导变元之一（即，在上述两个关键字之一或全部之后或者引导变元中不存在任何一个关键字）时，网络支持包 3220 将从 BOOTP 服务器检索寻址和其他引导参数，如方块 3316 所示。网络支持包 3220 将生成 / 创建引导串，如方块 3318 所示。然后，网络支持包 3220 使用默认的自举协议 TFTP 从远程服务器加载引导映像，如方块 3320 所示。然后，固件在计算机系统 3100 处完成引导过程，如方块 3322 所示。

[0197] 下面的例子提供了以下两者的说明：(a) 用于通过 IPv4 寻址格式和 BOOTP 发现协议完成引导映像检索的常规引导变元，以及 (b) 用于通过 IPv6 寻址格式和/或使用 DHCP 发现协议完成引导映像检索的扩展引导变元。后者的例子依赖于使用常规自举方法中提供的其他变元之外的特定关键字和变元。以下是前两个常规例子。

[0198] 例子 1：对于本例子，在命令行中提供所有变元。不使用引导发现协议。此外，所有地址（即 siaddr、ciaddr）均以 IPv4 格式表达。

[0199] open network-device:promiscuous, speed = 100, duplex = full, siaddr, filename, ciaddr, giaddr, bootp-retries, tftp-retries, subnet-mask

[0200] 例子 2：在第二个常规例子中，使用 BOOTP 服务器来获取 siaddr 和 filename。其他未指定的变元采用默认值。

[0201] open network-device:duplex = full, bootp

[0202] 现在，以下例子示出了用于选择备选引导发现协议和不同的地址表示（即 DHCP 和
IPv6) 的方法。
例 3：对于当前例子，使用由 diaddr 指定的 DHCP 服务器来获取 ciaddr、siaddr 和 filename。其他未指定的变量采用默认值。diaddr 以 IPv6 格式表达。
例 4：在本例子中，网络支持包发出广播查询以查找 DHCP 服务器。然后，使用由查询返回的 DHCP 服务器来获取 ciaddr、siaddr 和 filename。此外，其他未指定的变量采用默认值。diaddr 以 IPv4 格式表达。
例 5：对于本例子，使用 DHCP 服务器获取 ciaddr。其他未指定的变量采用默认值。
例 6：在此最终的例子中，使用 siaddr 和 diaddr 连接到引导服务器并获取引导文件的文件名。siaddr 和 diaddr 以 IPv6 格式表达。其他未指定的变量采用默认值。
例 7：在本例子中，用于使得 InfiniBand(IB) 服务器能够通过 IB 网络从远程存储设备引导操作系统。增强该设备的固件，以便在表示计算机设备的硬件的层次数据结构中提供 IB 主机通道适配器 (HCA) 的表示。相应地，定义新的设备类型，“hca”设备类型，其表示 IB HCA。hca 设备在层次固件数据结构内通过节点全局唯一标识符来描述特征。hca 设备含有用于开启和关闭设备的接口以及从设备读取和写入设备的接口。“hca”设备可以具有一个或多个“子设备”，其被称为“ib 端口”设备，并且由端口全局唯一标识符来标识。除了为父 hca 设备定义的开启、关闭、读和写接口以外，ib 端口设备还包括使得该设备能够被用作引导设备的加载接口。
例 8：提供了一种“ib- 引导”支持包，并且定义引导变元集以便由该 ib- 引导支持包利用。ib- 引导支持包接收并特定关键字一起指定目标 IB 端节点的引导变元。利用固件结构的层次特性，基于指定关键字来将这些 IB 专用变元传递至适当的支持包，以便完成引导处理。ib- 引导支持包还实现特定的分组传送协议，当通过 IB 网络进行引导时，通过 hca 设备驱动器来利用这些特定的分组传送协议。
例 9：在一个实施例中，提供了特定关键字，其标识 SCSI（小型计算机系统接口）RDMA（远程直接存储器访问）协议——SRP——支持包。SRP 支持包指定将要实现存储器引导方法，由此，利用 SRP 来通过 IB 结构传输 SCSI 命令以及访问引导设备。主机设备然后从通过通道适配器连接至 I/O 结构的存储设备（引导服务器）读取数据块（包括引导映像）。因而通过 SRP 支持包提供了对引导服务器的访问，并且经由 IB 网络将引导映像返回至主机设备用于完成主机设备的引导操作。
例 10：现参照附图，图 4A 根据本发明描述的实施例，提供了利用 InfiniBand(IB) 网络连接来与各种外围和组网设备（包括一个或多个远程引导服务器）进行通信的组网计算机系统的主要硬件组件的高级表示。如所示出的，计算机系统 4100（文中可互换地称为 IB 网络
内的主计算机或主计算机设备）包括经由存储总线 4103 耦合于存储器 4102 的中央处理器（CPU）4101，存储总线 4103 又通过输入/输出（I/O）总线接口 4105 提供给各种其它外围或网络设备的连接。CPU 4101 是执行来自主存储器 4102 的指令以及处理来自主存储器 4102 的数据的通用可编程处理器。主存储器 4102 可以是动态随机访问存储器（DRAM），其中，从存储器（4123/4124）加载数据用于由 CPU 4101 进行处理。

[0215] 存储总线 4103 提供了用于在 CPU 4101、主存储器 4102 和 I/O 总线接口 4105 之间传送数据的数据通信路径。I/O 总线接口 4105 进一步耦合于系统 I/O 总线 4104，其用于传送数据到各种 I/O 单元和传送来自各种 I/O 单元的数据。I/O 总线接口 4105 通过系统 I/O 总线 4104 与多个 I/O 接口单元 4111-4113 进行通信，I/O 接口单元 4111-4113 也被称为 I/O 处理器（1OP）或 I/O 适配器（IOA）。系统 I/O 总线 4104 可以是工业标准 PCI 总线或任何其它适当的总线技术。I/O 接口单元支持与各种存储器和 I/O 设备进行通信。举例来说，终端接口单元 4111 支持连接一个或多个用户端 4121-4122。存储接口单元 4112 支持连接一个或多个直接访问存储设备（DASD）4123-4124（其可以表示独立磁盘冗余阵列（RAID）存储器，被配置作为主计算机的单个大型存储设备）。各种其它类型的 I/O 设备（未示出）可以依附于 I/O 接口单元，例如磁带驱动、打印机和网络收发机。

[0216] 根据本发明的实施例，I/O 接口单元之一是 IB I/O 适配器 /HCA4113，由于 IB I/O 适配器 / HCA 4113 使得在一个计算机 4100 能够连接至 IB 网络，因此文中可互换地将 IB I/O 适配器 / HCA 4113 称为主机通道适配器（HCA）。HCA 是可编程 DMA（直接存储器访问）引擎，其生成和消耗分组。

[0217] IB I/O 适配器 4113 支持到 InfiniBand 网络 4130 的连接，InfiniBand 网络 4130 包括交换机 4131 和多个点到点连接 4132-4135。InfiniBand 网络是可以用于连接至各种设备的本地网。如上所说明的，InfiniBand 网络 4130 提供到其它计算机系统的连接，例如 IB 引导服务器 4142 经由目标通道适配器（TCA）4141 到交换机 4131。InfiniBand 网络 4130 还提供到路由器或适配器的连接，而路由器或适配器又将 InfiniBand 网络 4130 耦合到其它局域网或广域网。在图 4A 的示例性实施例中，TCA 4139 经由以太网适配器 4140 连接至局域网（LAN）4145。LAN 4145 提供对远程引导服务器 4150 的访问。此外，路由器 4138 连接至因特网 4146，其提供对第二远程引导服务器 4151 的访问。如文中所提到的，除了 InfiniBand I/O 适配器（HCA）4113 和相连的 IB 网络 4130 以外，文中将计算机系统 4100 的组件也称为“主计算机”，以便区别于来自跨 IB 网络 4130 的引导服务器的引导 OS 的主机设备。如所说明的，提供了两个外部网络，即 LAN4145 和因特网 4146。将通过主机设备 4100 对因特网 4146 和 LAN 4145 的访问描述为发生在 IB 网络 4130 的“端节点”，例如 TCA 4139。

[0218] 除了以上组件之外，IB 网络 4130 进一步包括连接到 SCSI 4162 的第三 TCA 4160。存储区域网（SAN）4164 连接到 SCSI 设备 4162，该存储区域网 4164 包括存储服务器 4166。存储服务器 4166 进一步包括一个或多个子设备，其在文中被称为逻辑单元（LUN）4167/4168，并且连接至存储服务器 4166。至少一个 LUN 4167/4168 含有被利用来引导主计算机 4100 的引导映像。如以下所描述的，可以通过盘支持包来访问存储在 LUN 4167/4168 上的包括引导映像在内的数据。因而，通过直接或经由存储区域网 4164，而依附于远程 IB 节点（第三 TCA 4160）的存储设备（LUN 4167/4168），支持远程引导操作。

[0219] 尽管图 4A 中所示的 IB 网络 4130 具有单个交换机 4131 和对于相应的数字设备的
五个连接 4132-4136，然而应当理解，IB 网络可以具有不同数目的交换机、适配器、路由器和链路（但可能不具有交换机或路由器）。IB 网络的拓扑可以是显著不同并且更加复杂的，并且因而不同数目的设备可以耦合于 IB 网络。

[0220] 应当理解，图 4A 旨在高级地描绘计算机系统 4100 的代表性主要组件，并且各个组件可以具有比图 4A 中所表示的组件更大的复杂性。应当进一步认识到，可以给出不同于或除了图 4A 中所示的那些组件以外的组件，并且这样的组件的数目、类型和配置是可以改变的。因而，仅仅通过举例的方式给出图 4A，且并不旨在隐含对本发明的任何限制。

[0221] 本发明通过跨 IB 网络 4130 发生的引导过程，为连接至 IB 网络 4130 的计算机设备 4100 启用了远程引导过程。本发明的一个实施例使得能够在 IB 结构上完成网络引导过程，即存储器引导方法。该存储器引导方法涉及连接至位于远程 IB 端节点的存储设备（例如，LUN 4167/4168）以及使用诸如 SRP 的传输协议来获得引导文件。

[0222] 如文中较为详细的描述的，在远程引导操作的全面实施中，通过计算机设备 4100 来支持特定的存储器引导方法。存储器引导方法涉及连接至直接或通过 SAN 4164 而依附于远程 IB 节点（第二 TCA 4160）的存储设备 (LUN4167/4168)，以及通过 IB 结构发送存储协议命令，例如 SCST（小型计算机系统接口）RDMA（远程直接存储器访问）协议 (SRP)。

[0223] 在以上示例计算机系统和网络环境中，本发明的第一方面涉及修改系统固件，以便使得该固件能够标识和表示在系统初始化期间系统固件按照使得 IB 端节点用于引导操作的方式所生成的层次结构内的 IB 端节点。在计算机系统通电（或重新引导）之后，系统经历引导过程，通过该引导过程，系统固件通过标识和建立 / 启动系统中的设备来准备用于操作的系统。在引导过程中，固件将系统硬件表示为具有附属设备的互连总的层次。该层次表示描述了用户配置选择，并且含有用于硬件设备的固件设备驱动器以及由那些驱动器使用的支持路由器。

[0224] 在 IB 网络内，增强主机系统的引导固件以便识别存在 IB HCA 以及一个或多个 IB 端节点设备。根据一个实施例，将新的设备类型定义为“hca”设备类型，其表示 IB HCA。hca 设备的特征在层次固件数据结构中通过节点全局唯一标识符来描述。hca 设备包含有开启和关闭设备的接口以及从设备读取和写入设备的接口。hca 设备包括若干“子设备”，其被称为“ib 端口”设备并且由端口全局唯一标识符来标识。除了为父 hca 设备定义的开启、关闭、读和写接口以外，ib 端口设备还包括使该设备能够作为引导设备来使用的加载接口。

[0225] 本发明的第二相关方面涉及命令关键字的创建和利用，该命令关键字被添加到在系统初始化期间由系统固件所生成的引导变元。在一个实施例中，命令关键字指定分组传送协议和关联的变元集。通过 ib- 引导支持包来理解这些命令关键字，ib- 引导支持包接收来自系统固件的引导变元。命令关键字通过指定待采用的具体网络引导方法来修改与自举方法关联的引导变元。根据所描述的实施例，命令关键字是引导变元中的第一变元，并且每次只可以指定一个命令关键字。如以下进一步描述的，在一个实施例中，所有的命令关键字都是互斥的，并且因而，仅将关键字之一添加到标准引导变元之前来启用 IB 级远程引导操作。如果指定了超过一个的命令关键字，则发送具有警告消息的失败结果给主计算机设备的用户并且该设备不开启。

[0226] 通过在用于 ib- 引导支持包的开启接口的变元中使用命令关键字来选择特定的网络引导方法。该命令关键字指定了分组传送协议和关联的变元集。在本说明书中还描
述了对于多个命令关键字的使用和 / 或应用。在本发明的一个实施例中，定义命令关键字“srp”，以便选择存储器引导方法，并且使用 SRP 提供了特定的引导功能。

[0227] srp 关键字指定了存储器引导方法，当指定存储器引导方法时，其触发使用 SCSI RDMA 协议 (SRP)，以便在 IB 结构上传输 SCSI 命令并访问引导设备（图 4A 的 LUN 4167/4168）。引导平台（即，主机设备）通过适配器连接至 IB 结构的存储设备读取数据块。当指定了存储器方法时，提供 IB 专用引导变元集，并且其后是与盘相关的变元。
通过 ib- 引导开启方法将与盘相关的变元传递至盘支持包。在由开放固件工作组发行的 PowerPC Micro-processor Common Hardware Reference Platform (CHRP) Binding 中描述了对盘变元的利用，在此通过引用的方式将其相关内容纳入本文。

[0228] 根据本发明的一个实施例，用于 SRP 引导的引导变元如下：

[0229] open ibport-device:iocguid = 10C-GUID, portgid = 10C-GID, srp, srpiport = init-portid, srptport = target-port-id, tlun = target-lun, disk-labelargs

[0230] 在以上引导变元中，对特定变元的描述包括：

[0231] (a) iocguid = 10C-GUID；其是标识目标 I/O 控制器或 I/O 单元的 64 位全局唯一 EUI-64 顺应标识符；

[0232] (b) portgid = 10C-GID；其是标识目标 I/O 单元中的端口的 128 位。以 IPv6 格式表示该变元并且该变元由 64 位 GID 前缀和分派给 I/O 控制器的 64 位端口 GUID 组成；

[0233] (c) srpiport = init-portid；其是唯一标识 SRP 启动器端口的 128 位。该数含有用于启动器端口的 GUID 和确保所有的 SRP 启动器端口标识符都是唯一的 64 位数；

[0234] (d) srptport = target-port-id；其是唯一标识 SRP 目标端口的 128 位。该数含有用于目标 I/O 控制器的 GUID 和目标端口的服务 ID；以及

[0235] (e) tlun = target-lun；其是 SRP 目标设备的 64 位逻辑单元号。

[0236] 如文中所提供的，10C-GUID 和 10C-GID 唯一标识 InfiniBand 结构中的端节点。ib- 引导过程建立与目标 I/O 控制器或端节点的连接，并且然后利用通过命令关键字指定的协议在 IB 传输上发送分组。在本说明书中还描述了网络相关的变元 (network-args) 的例子。

[0237] 本发明的第三方面涉及 ib- 引导支持包的扩展，以便能够无缝连接到来自远程引导设备的 SDP 网络引导过程或 IvoIB 网络引导过程的 SRP 网络引导过程。提供 ib- 引导支持包来协助封装使用 ib 端口设备在 IB 网络上进行引导的变元，并且定义引导变元集来由 ib- 引导支持包利用。ib- 引导支持包处理指定了目标 IB 端节点（即，耦合于 IB 结构的 TCA 或路由器或适配器，以及直接或经由网络连接到 IB 结构的引导服务器）的引导变元。在一个实施例中，ib- 引导支持包还实现通常由 hca 设备驱动器利用的特定分组传送协议，当在 IB 上引导时利用该特定分组传送协议。

[0238] 利用固件结构的层次特性，基于如 ib- 引导支持包所处理的指定关键字，将这些 IB 专用变元传递至适当的支持包，以便启动对引导映像的检索。因而，当提供给引导变元的关键字是 srp 时，ib- 引导支持包将引导变元传送至 SRP 支持包。可选地，并且如本说明书中所描述的，关键字可以指定使用网络支持包来完成网络引导方法，或者 ib- 引导支持包可以将引导变元传递至 SDP 支持包以完成 IB 网络引导。

[0239] 图 4B 是根据本发明的一个实施例描绘了一过程的流程图，通过该过程，系统固件
发现IB HCA并且表示层次拓扑内的设备及其子设备。参照图4A所示的系统组件描述了图4B的过程。过程从块4202开始，由此，自主机系统4100（图4A）的主机系统上，由诸如系统固件4145（图4A）的固件启动硬件初始化。如块4204所示，在系统初始化期间，系统固件4145遍历系统总线并且标识出主机系统4100的I/O拓扑。在判定块4206处，如果在附件所述的任何设备中检测到选选项ROM，则加载该选项ROM并且执行其代码，如块4208所示。作为执行代码的结果，创建设备的表示，并将其添加到主层次I/O拓扑，如块4214所示。

[0240]在判定块4206处，如果未检测到选选项ROM，则在块4210处系统固件4145确定是否检测到 IB HCA（或“hca”设备）。如果检测到 IB HCA设备，则系统固件4145创建hca设备及其子ib端口设备的表示，如块4214所示。系统固件4145还创建I/O拓扑内ib端口设备的位置的设备表示，如块4214处所指示的。在判定块4210处，如果未检测到 IB HCA，那么过程结束，如块4211所示的。因为“ib端口”设备是利用特定的引导变量重启动的，所以系统固件4145获得开启ib端口设备所需要的引导变量，如块4216所示。系统固件然后将引导变量传递至ib—引导支持包用于处理，如块4218所示的。

[0241]现至图4C，其根据本发明的一个说明性实施例，提供了示例系统的框图表示，该示例系统具有这样的主要组件（硬件和软件实用体），即其使得能够通过IB—引导支持包来利用IB—引导变量，从而启用来自远程引导设备的SRP引导或IBoIP引导或SDP引导。如所示出的，正被引导的设备的系统固件4145（图4A），例如，主机系统4100（图4A）在通信上耦合于ib—引导支持包4310。ib—引导支持包建立与引导变量中所指定的目标I/O控制器的连接。当从IB连接的引导设备引导时，ib—引导支持包4310传输SDP引导变量到SDP支持包4330，其利用SDP网络协议来获得来自引导服务器4137的引导文件。当通过通过IP网络连接的服务器引导时，ib—引导支持包4310将IPOiB—引导变量转发到增强的网络支持包4320。最后，当从连接到服务器（其又直接或经由SAN连接至TCA）的LUN（或其他设备）引导时，ib—引导支持包4310将SCSI引导变量转发到SRP支持包4340。

[0242]所示出的SRP支持包4340连接至存储服务器4166，存储服务器4166具有与其相连的LUN4167/4168。在服务器4166中的是盘支持包4255，利用盘支持包4255接收引导变量，访问LUN4167/4168中适当的一个，并且从LUN4167/4168之一检索引导映像/文件。然后将所检索的引导映像/文件经由IB网络提供给主机系统固件4145。

[0243]SRP支持包有助于在IB链路上发送SCSI命令，以便到达存储服务器，并且SRP支持包然后向盘支持包传送控制。盘支持包完成从远程SCSI盘（LUN4148）对引导映像的实际加载。

[0244]IP引导服务器4150和IB引导服务器4166能够将引导映像提供给计算机系统4100（图4A）的系统固件4145，以便使得计算机系统4100能够完成引导操作。服务器4150和4137分别从网络支持包4320和SDP支持包4330接收相应的引导变量。然而，网络支持包4320通过IPoIB消息接收变量并且传输IPoIB消息，这与SDP支持包4330所支持的消息传输的类型/格式（即IB信息）不同。

[0245]图4D描绘了在ib—引导支持包与根据包括在引导变量内的命令关键字所选择的其它三种支持包之间的引导变量的流程。ib—引导支持包提供指定了目标IB端节点的引导变量。具体而言，关于本发明，图4D的流程图描绘了当选择存储网络方法时，ib—引导包和SRP支持包之间的信息流。ib—引导包通过远程直接存储器访问（RDMA）接口向SRP支持
包传送控制以便促进 SCSI 命令的传输。

[0246] 利用固件结构的层次特性将这些 IB 专用变异传递至适当的支持包以便完成引导处理。过程开始于块 4402，其说明了 ib- 引导支持包从系统固件接收引导变元。ib- 引导支持包解析引导变元，如块 4404 所示，并且在块 4406 处确定变元是否指示 IPoIB 引导操作。如果指示 IPoIB 引导操作，则 ib- 引导支持包创建用于 IPoIB 引导的引导串（boot string），如块 4408 处所提供的，并且然后将创建的引导串转发至网络支持包 4320，如块 4410 所示。因而，如果选择了 IBoIP 网络方法，则 ib- 引导包向网络支持包 4320 传送控制，并且通过 ib- 引导支持包促进整个网络结构。

[0247] 如果引导变元未指示 IPoIB 引导，那么在块 4412 处进行引导变元是否指示 SDP 引导的下一定。如果指示 SDP 引导，则 ib- 引导支持包再次生成 SDP 引导所必需的引导串，如块 4414 所示，并且将生成的引导串转发至 SDP 引导支持包，如块 4416 处所描绘的。因而，当选择 IB 网络方法时，ib- 引导支持包向 SDP 支持包或在 InfiniBand 链路上促进或提供直接文件传输协议的任何其它的包传送控制。

[0248] 否则，如果引导变元既未指示 IPoIB 引导，也未指示 SDP 引导，则在块 4418 处进行是否指示 SRP 引导的最后确定。当指示 SRP 引导时，ib- 引导支持包生成 SRP 专用引导变元，如块 4420 所示，并且 ib- 引导支持包将引导串传送至 SRP 支持包 4340 以完成存储器引导过程，如块 4422 处所指示的。当未指示特定的引导方法时（或者如果指示了多种引导方法），则报告错误，如块 4424 所示，并且引导过程结束。

[0249] 在本说明书中还详细描述了网络支持包和 SDP 支持包接收和处理引导变元的过程。

[0250] 图 4C 是根据本发明的一个实施例描绘了对引导变元的处理的流程图。当利用在 IB 链路（IPoIB）上传输 TCP/IP 消息来加载来自 LAN 服务器（例如图 4A 的引导服务器 4150）的引导文件时，SRP 支持包 4340（图 4C）进行对引导变元的该处理。通过 I/O 拓扑表示的层次特性，由 SRP 支持包所进行的读和写操作被传递至 ib- 引导支持包并且实际上是由 HCA 设备实现的。HCA 设备封装 SCSI 命令并且经由 IB 链路将命令发送至指定的 IB 目标节点。参照图 4C 的组件进一步描述该过程。

[0251] 该过程开始于块 4502，其说明了 SRP 支持包 4340（图 4C）从 ib- 引导支持包 4310（图 4C）接收 ib- 引导变元。SRP 支持包 4340 解析引导变元，如块 4504 所示，并且然后通过 IB 将 SCSI 命令发送至存储服务器 4164（图 4C），如块 4506 处所指示的。在块 4508 处确定是否已经完成对 SCSI 命令的传输，并且当已经完成传输时，在块 4510 处，将特定的加载变元传递至盘支持包 4255（图 4C），并且盘支持包 4255 导致启动引导过程来从目标盘（LUN 4168）引导映像文件。目标设备（4167/4168）提供引导映像，在主机计算机 4100 上加载引导映像，如块 4512 所示。然后，在主计算机 4100（图 4A）处着手并完成引导操作，如块 4514 所指示的。

[0252] 本发明的附图的描述涉及特定术语 / 专门名词的使用，其中某些由 IETF 标准部门所定义。当其名称被用于描述类似的功能和 / 或功能性时，应当理解，本发明在范围内扩展以覆盖所有一般功能和 / 或特征，而不被所使用的命名法。还应当理解，特定的参数名字的使用仅是用于举例而不是意味着暗示对本发明的任何限制。因此，本发明不受任何限制地通过被用于描述上面的参数不同的命名法 / 专门名词来实现。
作为最后一个问题，重要的是：尽管在安装有软件的全功能计算机系统的环境中，已经并将继续描述本发明的说明性实施例，但是本领域的技术人员将意识到，本发明的说明性实施例的软件方面能够以各种形式作为程序产品分布，并且无论被用于实际实现该分布的特定类型的信号承载媒体如何，本发明的说明性实施例都等地适用。信号承载媒体的例子包括可记录类型媒体（诸如软盘、硬盘驱动器、CD ROM）和诸如数字和模拟通信链路之类的传输类型媒体。

尽管已经参考优选的实施例具体地示出并描述了本发明，但是本领域的技术人员应当理解，在不偏离本发明的精神和范围的条件下可以在形式和细节中做出各种改变。
图 1A
图 1D

开始硬件初始化

遍历总线

可选 ROM？

是 加载并执行可选 ROM 代码

否 ISCSI TOE？

是 创建 "iSCSI" 设备

否 创建 "网络" 设备

网络适配器？

是

将设备添加到分层 结构

处理网络参数（模式、模式）

将参数传递给网络支持包

结束
图 1E
图 2A
图 2 B
图 2E
图 3A
图 3 B
图 3C
图 4C
图 4D
图 4E