(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112671663 B
(45)授权公告日 2024.12.20

(21)申请号 202011371223.0
(22)申请日 2017.03.02
(65)同一申请的已公布的文献号
申请公布号 CN 112671663 A
(43)申请公布日 2021.04.16
(30)优先权数据
15/088948 2016.04.01 US
(62)分案原申请数据
201780014783.0 2017.03.02
(73)专利权人 英特尔公司
地址 美国加利福尼亚州
(72)发明人 F.G.伯纳特 K.库马尔
T.威尔哈尔姆 R.K.拉马努詹
B.斯莱克塔
(74)专利代理机构 中国专利代理(香港)有限公司 72001
专利代理师 付曼 姜冰
(51)Int.Cl.
H04L 49/15(2022.01)
H04L 49/103(2022.01)
H04L 47/125(2022.01)
H04L 47/24(2022.01)
H04L 49/20(2022.01)
(56)对比文件
US 2013290967 A1,2013.10.31
审查员 赖思

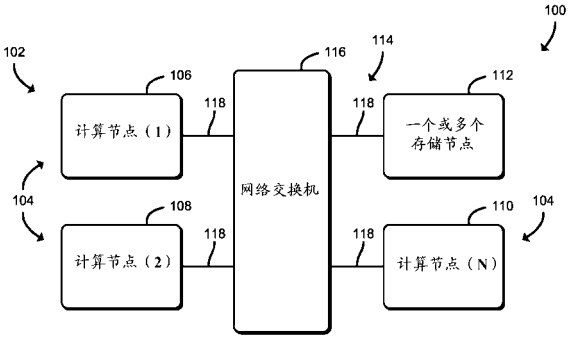
权利要求书3页 说明书21页 附图7页

(54)发明名称

结构架构中用于基于服务质量进行节流的方法和设备

(57)摘要

用于在结构架构中基于服务质量进行节流的技术包括经由互连结构跨结构架构互连的多个网络节点的网络节点。该网络节点包括主机结构接口(HFI),其被配置成促进往/来于网络节点的数据传送,监视用于处理和传送数据的网络节点的资源的服务质量等级,基于监视的服务质量等级的结果检测节流条件。HFI还被配置成响应于已经检测到节流条件,生成节流消息并将其传送到互连网络节点中的一个或多个。HFI另外被配置成从网络节点中的另一个网络节点接收节流消息,并基于接收到的节流消息对资源中的一个或多个执行节流动作。本文描述了其他实施例。



CN 112671663 B

1.一种设备,包括:

第一多芯片封装,所述第一多芯片封装包括:

第一多个核;

与所述第一多个核耦合的第一互连;

用于传送数据的互连链路;

经由所述互连链路耦合到所述第一互连的第二互连;

第一存储器互连,用于将所述第一多个核耦合到第一系统存储器装置,所述第一多个核经由所述第一互连、所述互连链路和所述第二互连来访问所述第一存储器互连;

其中,所述第一存储器互连和第一系统存储器装置要与第一非统一存储器访问NUMA域相关联;

其中,所述第一多芯片封装耦合到与第二NUMA域相关联的第二多芯片封装;

其中,第一NUMA域标识符要与所述第一NUMA域相关联,并且第二NUMA域标识符要与所述第二NUMA域相关联;

用于监视与所述第一NUMA域相关联的资源的利用的监视电路,所述监视电路包括一个或多个模型特定寄存器MSR,以存储与访问所述资源的请求相关联的计数器值,所述计数器值包括与来自所述第一NUMA域的所述资源的利用相关联的第一计数器值以及与来自至少所述第二NUMA域的所述资源的利用相关联的第二计数器值;以及

执行电路,用于根据所述计数器值中的一个或多个来限制来自所述第一NUMA域或者来自所述第二NUMA域的所述资源的利用。

2.如权利要求1所述的设备,其中,所述第一多芯片封装还包括处理器,所述处理器包括所述第一多个核。

3.如权利要求2所述的设备,其中,所述处理器还包括多个高速缓存级,所述多个高速缓存级包括第1级,即L1,高速缓存。

4.如权利要求1至3中任一项所述的设备,还包括要在所述第一多芯片封装和所述第二多芯片封装之间形成的虚拟信道,所述虚拟信道携带节流消息以指示对所述第一多个核的一个或多个执行资源进行节流的节流请求。

5.如权利要求1至3中任一项所述的设备,还包括用于将所述第一多个核耦合到一个或多个装置的I/O接口。

6.如权利要求1至3中任一项所述的设备,其中,所述第一NUMA域与第一套接字相关联,并且所述第二NUMA域与第二套接字相关联。

7.如权利要求1至3中任一项所述的设备,其中,所述互连链路要根据一致性协议来传送数据。

8.一种方法,包括:

提供第一多芯片封装,所述第一多芯片封装包括第一多个核、与所述第一多个核耦合的第一互连、用于传送数据的互连链路、经由所述互连链路耦合到所述第一互连的第二互连、用于将所述第一多个核耦合到第一系统存储器装置的第一存储器互连,所述第一多个核经由所述第一互连、所述互连链路和所述第二互连来访问所述第一存储器互连;

将所述第一存储器互连和第一系统存储器装置与第一非统一存储器访问NUMA域相关联,其中第一多芯片封装耦合到与第二NUMA域相关联的第二多芯片封装;

将第一NUMA域标识符与所述第一NUMA域相关联;

将第二NUMA域标识符与所述第二NUMA域相关联;

监视与所述第一NUMA域相关联的资源的利用,包括使用一个或多个模型特定寄存器MSR,以存储与访问所述资源的请求相关联的计数器值,所述计数器值包括与来自所述第一NUMA域的所述资源的利用相关联的第一计数器值以及与来自至少所述第二NUMA域的所述资源的利用相关联的第二计数器值;以及

根据所述计数器值中的一个或多个来限制来自所述第一NUMA域或者来自所述第二NUMA域的所述资源的利用。

9.如权利要求8所述的方法,其中,所述第一多芯片封装还包括处理器,所述处理器包括所述第一多个核。

10.如权利要求9所述的方法,其中,所述处理器还包括多个高速缓存级,所述多个高速缓存级包括第1级,即L1,高速缓存。

11.如权利要求8至10中任一项所述的方法,还包括:在所述第一多芯片封装和所述第二多芯片封装之间形成虚拟信道,所述虚拟信道携带节流消息以指示对所述第一多个核的一个或多个执行资源进行节流的节流请求。

12.如权利要求8至10中任一项所述的方法,还包括:经由I/O接口将所述第一多个核耦合到一个或多个装置。

13.如权利要求8至10中任一项所述的方法,还包括:将所述第一NUMA域与第一套接字相关联,并且将所述第二NUMA域与第二套接字相关联。

14.如权利要求8至10中任一项所述的方法,还包括:根据一致性协议经由所述互连链路来传送数据。

15.一种设备,包括:

用于提供第一多芯片封装的部件,所述第一多芯片封装包括第一多个核、与所述第一多个核耦合的第一互连、用于传送数据的互连链路、经由所述互连链路耦合到所述第一互连的第二互连、用于将所述第一多个核耦合到第一系统存储器装置的第一存储器互连,所述第一多个核经由所述第一互连、所述互连链路和所述第二互连来访问所述第一存储器互连;

用于将所述第一存储器互连和第一系统存储器装置与第一非统一存储器访问NUMA域相关联的部件,其中第一多芯片封装耦合到与第二NUMA域相关联的第二多芯片封装;

用于将第一NUMA域标识符与所述第一NUMA域相关联的部件;

用于将第二NUMA域标识符与所述第二NUMA域相关联的部件;

用于监视与所述第一NUMA域相关联的资源的利用的部件,所述监视包括使用一个或多个模型特定寄存器MSR,以存储与访问所述资源的请求相关联的计数器值,所述计数器值包括与来自所述第一NUMA域的所述资源的利用相关联的第一计数器值以及与来自至少所述第二NUMA域的所述资源的利用相关联的第二计数器值;以及

用于根据所述计数器值中的一个或多个来限制来自所述第一NUMA域或者来自所述第二NUMA域的所述资源的利用的部件。

16.如权利要求15所述的设备,其中,所述第一多芯片封装还包括处理器,所述处理器包括所述第一多个核。

17.如权利要求16所述的设备,其中,所述处理器还包括多个高速缓存级,所述多个高速缓存级包括第1级,即L1,高速缓存。

18.如权利要求15至17中任一项所述的设备,还包括:用于在所述第一多芯片封装和所述第二多芯片封装之间形成虚拟信道的部件,所述虚拟信道携带节流消息以指示对所述第一多个核的一个或多个执行资源进行节流的节流请求。

19.如权利要求15至17中任一项所述的设备,还包括:用于经由I/O接口将所述第一多个核耦合到一个或多个装置的部件。

20.如权利要求15至17中任一项所述的设备,还包括:用于将所述第一NUMA域与第一套接字相关联的部件,以及用于将所述第二NUMA域与第二套接字相关联的部件。

21.如权利要求15至17中任一项所述的设备,还包括:用于根据一致性协议经由所述互连链路来传送数据的部件。

22.一种机器可读介质,其上面存储指令,所述指令在被执行时导致所述机器执行如权利要求8-14中任一项所述的方法。

结构架构中用于基于服务质量进行节流的方法和设备

[0001] 相关美国申请的交叉引用

[0002] 本申请要求2016年4月1日提交的、名为“TECHNOLOGIES FOR QUALITY OF SERVICE BASED THROTTLING IN FABRIC ARCHITECTURES”的、序列号为15/088,948的美国实用专利申请的优先权。

背景技术

[0003] 由个体、研究人员和企业对计算装置的增加的计算性能和存储容量的需求已经导致被开发来满足这些需求的各种计算技术。例如,诸如基于企业云的应用(例如,软件即服务(SaaS)应用)、数据挖掘应用、数据驱动建模应用、科学计算问题解决应用等的计算密集型应用通常依赖于复杂的、大规模计算环境(例如,高性能计算(HPC)环境、云计算环境等)以执行计算密集型应用,以及存储大量数据。此类大规模计算环境可以包括经由高速互连(例如,采用统一结构的互连结构)连接的数千个(例如,企业系统)到数万个(例如,HPC系统)的多处理器/多核网络节点。

[0004] 为了实行此类处理器密集型计算,已经实现了各种计算技术以跨不同的网络计算装置分配工作负载,例如并行计算,分布式计算等。以此类分布式工作负载的操作的支持,多处理器硬件架构(例如,共享存储器的多个多核处理器)已被开发来使用各种并行计算机存储器设计架构来促进跨本地和远程共享存储器系统的多处理(即,由多个处理器来协调的、同时处理),并行计算机存储器设计架构诸如非统一存储器访问(NUMA)、和其他分布式存储器架构。

[0005] 因此,来自多个互连网络节点的存储器请求可以与具体网络节点的本地存储器请求占用相同共享缓冲器(例如,超级队列、请求表等)。然而,此类共享的缓冲器在大小上是有限的(例如,包含数十个条目),这可能导致其他存储器请求被排队,直到对于当前在共享缓冲器中的那些存储器请求的数据从存储器子系统返回。像这样,共享缓冲器的条目倾向于被那些针对提供高时延访问的存储器的那些存储器请求(例如,从远程网络节点接收的存储器请求)或正被过度利用的存储器的那些存储器请求占用。因此,由于没有可用于执行所述存储器请求的可用共享缓冲器条目,因此针对更快或非拥塞存储器的其他请求(例如,本地存储器请求)(即,将更快地被服务的存储器请求)可能在核中变得饥饿。

附图说明

[0006] 本文描述的概念在随附附图中作为示例而不是作为限制被示出。为了示出的简单和清晰,附图中示出的元件不一定按比例绘制。在认为合适的情况下,引用标记在附图中已经被重复以指示相应或类似的元件。

[0007] 图1是用于在结构架构中基于服务质量进行节流的系统的至少一个实施例的简化框图,该系统包括经由互连结构通信地耦合的多个互连网络节点;

[0008] 图2是图1的系统的网络节点之一的至少一个实施例的简化框图;

[0009] 图3是图2的网络节点的另一实施例的简化框图;

[0010] 图4是可以由图2的网络节点建立的环境的至少一个实施例的简化框图；

[0011] 图5是可以由图2的网络节点来执行的用于处理来自远程网络节点的本地存储器请求的方法的至少一个实施例的简化流程图；

[0012] 图6是可以由图2的网络节点来执行的用于访问远程网络节点的存储器的方法的至少一个实施例的简化流程图；

[0013] 图7是可以由图2的网络节点来执行的用于生成对一个或多个远程网络节点的外部传输的节流消息的方法的至少一个实施例的简化流程图；以及

[0014] 图8是可以由图2的网络节点来执行的用于处理从远程网络节点接收的节流消息的方法的至少一个实施例的简化流程图。

具体实施方式

[0015] 虽然本发明的概念易受各种修改和替换形式的影响，但是其特定实施例已经在附图中作为示例被示出并且将在本文中详细描述。然而，应该理解的是，没有将本发明的概念限制于所公开的具体形式的意图，而是相反，意图在于覆盖与本发明和所附的权利要求书一致的所有修改、等同物和替代方案。

[0016] 在本说明书中对“一个实施例”、“一实施例”，“说明性实施例”等的引用指示所描述的实施例可以包括具体特征、结构或特性，但是每个实施例可以包括或可以不一定包括该具体特征、结构或特性。另外，此类短语不一定指的是相同实施例。此外，当结合一实施例描述具体特征、结构或特性时，无论是否明确描述，认为与其他的实施例结合来实现此类特征、结构或特性在本领域技术人员的知识之内。另外，应当领会的是，包括在以“A、B和C中的至少一个”的形式的列表中的项可以表示(A)；(B)；(C)；(A和B)；(A和C)；(B和C)；或(A、B和C)。类似地，以“A、B或C中的至少一个”的形式列示的项可以表示(A)；(B)；(C)；(A和B)；(A和C)；(B和C)；或(A、B和C)。

[0017] 在一些情况下，公开的实施例可以采用硬件、固件、软件或其任何的组合来实现。公开的实施例还可以被实现为由一个或多个暂态或非暂态机器可读(例如，计算机可读)存储介质(例如，存储器、数据存储等)携带的或存储在其上的指令，其可以由一个或多个处理器来读取和执行。机器可读存储介质可以实现为用于以由机器可读的形式存储或传送信息的任何存储装置、机构、或其他物理结构(例如，易失性或非易失性存储器、媒体盘、或其他媒体装置)。

[0018] 在附图中，一些结构或方法特征以特定布置和/或排序被示出。然而，应该领会的是，可能此类特定布置和/或排序可以不是必需的。而是，在一些实施例中，这些特征可以采用与附图中所示的方式和/或排序不同的方式和/或排序来布置。另外，在具体图中结构或方法特征的包括不意味着暗示在所有实施例中都需要此类特征，并且在一些实施例中，可以不包括此类特征或者此类特征可以与其他特征组合。

[0019] 现在参考图1，在说明性实施例中，用于在结构架构中基于服务质量进行节流的系统100包括经由互连结构114通信地耦合的多个互连的网络节点102。说明性系统100包括各种类型的网络节点102，包括多个计算节点104和存储节点112。说明性计算节点104包括被指定为计算节点(1)106的第一计算节点、被指定为计算节点(2)108的第二计算节点、以及被指定为计算节点(N)110的第三计算节点(即，计算节点104的“第N个”计算节点，其中“N”

是正整数并且指定一个或多个附加计算节点104)。应当领会的是,在其他实施例中,可以存在任何数量的计算节点104和/或存储节点112。说明性地,互连结构114包括用于通信地耦合网络节点102的网络交换机116和多个结构互连118。然而,应当领会的是,虽然仅示出了单个网络交换机116,但是在其他互连结构实施例中可以存在任何数量的网络交换机116。

[0020] 在使用中,网络节点102监视与本地资源(例如,物理和/或虚拟组件)相关联的服务质量等级,以检测与此类资源相关联的节流条件(例如,拥塞、饱和、过度利用、工作负载分配不公平性等)并在检测到这种节流条件时,将节流消息传送到结构架构的其他网络节点102,请求由接收网络节点102执行的节流动作。节流消息可以包括针对对网络节点102的具体资源进行节流的各种类型的节流请求。例如,节流消息可以包括存储器节流请求、I/O节流请求、加速器节流处理请求、HFI饱和节流请求等。应当领会的是,在检测到节流条件的时间段内会周期性地传送节流消息。换句话说,网络节点102继续传送节流消息直到相应的节流条件消失为止。

[0021] 为了这样做,不像在当前技术中网络节点102不外部地传送节流消息,从而使得节流局限于只能够节流对网络节点102本地的那些资源,网络节点102和相关联的互连结构114的组件被扩展成将节流信息传送(例如,生成新的节流消息、传播现有的节流信号等)到当前正请求对网络节点102的已经检测到节流条件的相应一个网络节点的共享结构(例如,共享缓冲器)的访问的其他网络节点102。

[0022] 在说明性示例中,诸如英特尔®一致性协议的高速缓存代理和归属代理的某些一致性协议包括代理实体被配置成将事务发起到一致存储器中(例如,经由高速缓存代理)并服务一致事务(例如,经由归属代理)。此类代理实体当前被配置成检测对网络节点102中的相应一个网络节点本地的某些条件,并发出本地处理器核节流信号来节流处理器的一个或多个核。然而,结构架构中的争用可以不仅发生在每个网络节点102内的共享路径中,而且在互连结构114的共享路径中,诸如共享缓冲器(例如,处理器核中的超级队列、高速缓存/归属代理中的请求表等)。

[0023] 在说明性示例中,计算节点(1)106可以正访问计算节点(2)108的存储器,其可以被配置成监视存储器访问请求(例如,本地接收的存储器访问、从计算节点104中的另一个计算节点接收的存储器访问等)和存储器利用等级。在某些条件下,计算节点(2)108可能由于存储器请求队列条目被已经从计算节点(1)106接收到的对计算节点(2)108的较慢存储器(例如,非高速缓冲存储器)的请求占用而经历高且不平等的争用。因此,在此类条件下,计算节点(2)108被配置成将节流消息传送到计算节点(1)106,指示计算节点(2)108的该存储器当前是饱和的,计算节点(1)可以使用该节流消息来降低指引向计算节点(2)108的存储器请求的注入速率。

[0024] 在一些实施例中,网络节点102被配置成使用开放系统互连(OSI)模型的传输层(即,层4(L4))来公开系统100的不同网络节点102之间的当前节点节流技术。因此,发源于网络节点102之一(例如,来自高速缓存代理、归属代理、输入/输出操作、调度器等)的新的和/或现有的节流信号可以通过结构互连118被传播到其他网络节点102,诸如请求对网络节点102的所述网络节点(所述节流信号发源于该网络节点)的共享结构的访问的那些网络节点。

[0025] 网络节点102可以实施为能够执行本文描述的功能的任何类型的网络业务(例如,

网络分组、消息、数据等)计算和/或存储计算装置,诸如但不限于,服务器(例如,独立式、机架安装式、刀片式等)、网络设备(例如,物理的或虚拟的)、交换机(例如,机架安装式、独立式、完全管理、部分管理、全双工、和/或半双工通信模式启用等),路由器、网页设备、分布式计算系统、和/或基于多处理器的系统。如前所述,说明性网络节点102包括计算节点104和存储节点112;然而,应当领会的是,网络节点102可以包括附加的和/或备选的网络节点,诸如控制器节点、网络节点、工具节点等,其为了保持描述的清晰而没有示出。

[0026] 如图2所示,说明性网络节点102包括指定为处理器(1)202的第一处理器、指定为处理器(2)208的第二处理器、输入/输出(I/O)子系统214、主存储器216、数据存储装置218、和通信电路220。应当领会的是,图1的计算节点104和/或存储节点112可以包括说明性网络节点102的图2中描述的组件。

[0027] 当然,在其他实施例中,网络节点102可以包括其他或附加组件,诸如一般在计算装置中发现的那些。另外,在一些实施例中,说明性组件中的一个或多个可以合并到另一组件中、或以别的方式形成其一部分。例如,在一些实施例中,高速缓冲存储器206或其部分可以合并在处理器202、208中的一个或两者中。此外,在一些实施例中,可以从网络节点102省略说明性组件中的一个或多个。例如,尽管说明性网络节点102包括两个处理器202、208,但是在其他实施例中,网络节点102可以包括更大量的处理器。

[0028] 处理器202、208(即,物理处理器封装)中的每一个可以实施为能够执行本文描述的功能的任何类型的多核处理器,诸如但不限于单个物理多处理器核芯片、或封装。说明性处理器(1)202包括多个处理器核204,而说明性处理器(2)208类似地包括多个处理器核210。如前所述,处理器202,208中的每一个包括多于一个的处理器核(例如,2个处理器核、4个处理器核、8个处理器核、16个处理器核等。)

[0029] 处理器核204、210中的每一个被实施为能够执行编程的指令的独立逻辑执行单元。在一些实施例中,处理器核204、210可以包括高速缓冲存储器的一部分(例如,L1高速缓存)和可用于独立地执行程序或线程的功能单元。应当领会的是,在网络节点102的一些实施例中(例如超级计算机),网络节点102可以包括数千个处理器核。处理器202、208中的每一个可以连接到被配置成接受单个物理处理器封装(即,多核物理集成电路)的网络节点102的母板(未示出)上的物理连接器或插口。

[0030] 说明性处理器(1)202另外包括高速缓冲存储器206。类似地,说明性处理器(2)208还包括高速缓冲存储器212。每个高速缓冲存储器206、212可以实施为与访问主存储器216相比相应的处理器202、208可以更快地访问的任何类型的高速缓冲存储器,例如管芯上的高速缓存或处理器上的高速缓存。在其他实施例中,高速缓冲存储器206、212可以是管芯外高速缓存,但是驻留在与相应处理器202、208相同的片上系统(SoC)上。应当领会的是,在一些实施例中,高速缓存存储器206、212可以具有多级架构。换句话说,在此类多级架构实施例中,高速缓冲存储器206、212可以实施为例如L1、L2或L3高速缓存。

[0031] 主存储器216可以实施为能够执行本文描述的功能的任何类型的易失性或非易失性存储器或数据存储装置。在操作中,主存储器216可以存储在网络节点102的操作期间使用的各种数据和软件,诸如操作系统、应用、程序、库、和驱动。主存储器216经由I/O子系统214通信地耦合到处理器202、208,I/O子系统214可以实施为用于促进与处理器202、208、主存储器216和网络节点102的其他组件的输入/输出的操作的电路和/或组件。例如,I/O子系

统214可以实施为或以别的方式包括存储器控制器中枢、输入/输出控制中枢、固件装置、通信链路(即,点对点链路、总线链路、导线、线缆、光导、印刷电路板迹线等)和/或用于促进输入/输出的操作的其他组件和子系统。在一些实施例中,I/O子系统214可以形成SoC的一部分并且与处理器202、208中的一个或两者、主存储器216、和/或网络节点102的其他组件一起被合并在单个集成电路芯片上。

[0032] 数据存储装置218可以实施为被配置用于数据的短期或长期存储的任何类型的装置或多个装置,诸如例如存储器装置和电路、存储卡、硬盘驱动器、固态驱动器、或其他数据存储装置。应当领会的是,数据存储装置218和/或主存储器216(例如,计算机可读存储介质)可以存储如本文所述的各种数据,包括能够由网络节点102的处理器(例如,处理器202、处理器208等)来执行的操作系统、应用、程序、库、驱动、指令等。

[0033] 通信电路220可以实施为能够实现通过网络的在网络节点102与其他计算装置(例如,计算节点104、存储节点112等)之间的通信的任何通信电路,装置或其集合。通信电路220可以被配置成使用任何一种或多种通信技术(例如,无线或有线通信技术)和相关协议(例如,因特网协议(IP)、以太网、蓝牙®,Wi-Fi®、WiMAX、LTE、5G等)来实现此类通信。

[0034] 说明性通信电路220包括主机结构接口(HFI)222。HFI 222可以实施为一个或多个内装板、子卡、网络接口卡、控制器芯片、芯片集或可由网络节点102使用的其他装置。例如,在一些实施例中,HFI 222可以与处理器202、208中的一个或两者集成(例如,在处理器202,208中的一个或两者内的一致结构上),实施为通过扩展总线(例如,高速PCI(PCIe))耦合到I/O子系统214的扩展卡、包括一个或多个处理器的SoC的一部分、或者被包括在也包含一个或多个处理器的多芯片封装上。另外或可选地,在一些实施例中,HFI 222的功能性可以在板级、插口级、芯片级和/或其他等级上被集成到网络节点102的一个或多个组件中。HFI 222被配置成促进向数据/消息的传送以使得在处理器202,208上执行的任务能够访问其他网络节点102的共享结构(例如,共享物理存储器),诸如在并行或分布式计算操作期间可能是必需的。

[0035] 应当领会的是,实现为存储节点112的那些网络节点102通常可以比实现为计算节点104的那些网络节点102包括更多的数据存储装置218容量。类似地,还应当领会的是,实现为计算节点104的那些网络节点102通常可以比实现为存储节点112的那些网络节点102包括更多处理器能力。换句话说,存储节点112可以实施为包括与计算节点104的存储装置的数量有关的多个硬盘驱动器(HDD)或固态驱动器(SDD)的物理服务器,而计算节点104可以实施为包括与存储节点112的处理器的数量有关的具有多个核的多个处理器的物理服务器。然而,应该进一步领会的是,无论与其他网络节点102有关的组件的配置如何,任何网络节点102都可以实施为计算节点104和/或存储节点112。

[0036] 再次参考图1,互连结构114(说明性地是网络交换机116和结构互连118的组合)可以实施为一个或多个总线、交换机和/或网络,其被配置成按各种互连协议和/或网络协议的功能来支持网络业务的传送。在使用中,互连结构114被网络节点102(例如,经由相应的HFI 222)用于与其他网络节点102(即,跨互连结构114)通信。因此,网络交换机116可以实施为能够在交换的或进行交换的结构架构中经由结构互连118进行网络业务转发的任何类型的交换装置(例如,纵横交换机)。

[0037] 现在参考在图3,在说明性实施例中,图2的网络节点102包括通信地耦合到HFI

222的一个或多个非统一存储器访问(NUMA)域300。说明性NUMA域300包括指定为NUMA域(1)302的第一NUMA域和指定为NUMA域(2)308的第二NUMA域。NUMA域300中的每个域包括物理处理器封装的多个分配的处理器核,所述物理处理器封装本文称为处理器。如说明性实施例中所示,NUMA域(1)302包括处理器(1)202的处理器核204并且NUMA域(2)308包括处理器(2)208的处理器核210。然而,应该领会的是,在一些实施例中,处理器202的处理器核204和/或处理器208的处理器核210可以被划分,并且每组划分的处理器核可以被分配到不同的NUMA域300。应该领会的是,分配到NUMA域300中的相应一个域的每组分配的处理器核可以被称为套接字核。换句话说,物理处理器封装的分配核的数量可以称为套接字。

[0038] 另外,每个NUMA域300对应于一具体存储器类型(例如,双倍数据速率(DDR)存储器、磁盘等),并且包括该存储器类型的本地存储器的一部分(例如,主存储器216),其已经被分配到相应NUMA域300的处理器核。此外,本地存储器直接链接到处理器核驻留所在的物理处理器封装。在说明性实施例中,NUMA域(1)302包括本地存储器(1)304,并且NUMA域(2)308包括本地存储器(2)310。在一些实施例中,可以经由互连314(例如,英特尔®超路径互连(UPI)、英特尔®快速路径互连(QPI)、AMD®统一媒体接口(UMI)互连、或诸如此类)在NUMA域300之间传送数据。NUMA域300之一的本地存储器被认为相对于其他NUMA域300是远程的或外来的。因此,应当领会的是,与使用本地存储器处理数据相比,跨互连314传送的网络业务可能引入负载/争用,增加总带宽使用,并降低与对远程存储器的访问相关的时延。

[0039] 每个说明性处理器202、208另外包括管芯上的互连(例如,处理器202的管芯上互连306和处理器208的管芯上互连312),其被配置成经由点对点接口316与HFI 222对接,所述管芯上的互连能够促进HFI 222与处理器202、208之间的数据传送。在一些实施例中,NUMA域300可以内部地被定义在HFI 222中。在说明性示例中,网络节点102之一(例如,计算节点(1)106)的NUMA域300之一(例如,NUMA域(1)302)可以对应于来自另一个网络节点102(例如,计算节点(2)108)的由HFI 222处理的事务。因此,计算节点(1)106的HFI 222可以在由计算节点(1)106确定计算节点(2)108正向计算节点(1)106传送太多请求时将节流消息发送到计算节点(2)108。在一些实施例中,此类节流消息可以包括经由点对点接口316由HFI 222接收的从处理器202的高速缓存代理传播的信息。

[0040] 现在参考图4,在说明性实施例中,网络节点102之一在操作期间建立的环境400。说明性环境400包括通信管理模块410、服务质量(QoS)监视模块420、节流消息传送模块430、节流消息接收模块440、和节流响应执行模块450。环境400的各种模块可以实施为硬件、固件、软件或其组合。照此,在一些实施例中,环境400的模块中的一个或多个可以实施为电气装置的电路或集合(例如,通信管理电路410、QoS监视电路420、节流消息传送电路430、节流消息接收电路440、节流响应执行电路450等)。

[0041] 应当领会的是,在此类实施例中,通信管理电路410、QoS监视电路420、节流消息传送输电路430、节流消息接收电路440和节流响应执行电路450中的一个或多个可以形成以下各项的一部分:一个或多个处理器(例如,图2的处理器(1)202和处理器(2)208)、I/O子系统214、通信电路220和/或网络节点102的其他组件。另外,在一些实施例中,说明性模块中的一个或多个可以形成另一模块的一部分和/或说明性模块中的一个或多个可以彼此独立。此外,在一些实施例中,环境400的模块中的一个或多个可以实施为虚拟化硬件组件或仿真架构,其可以由网络节点102的一个或多个处理器和/或其他组件建立和维护。

[0042] 在说明性环境400中,网络节点102进一步包括网络节点数据402、监视结果数据404、请求监视数据406、和NUMA标识数据408,其每个可以被存储在网络节点102的主存储器216中和/或数据存储装置218中。此外,网络节点数据402、监视结果数据404、请求监视数据406、和NUMA标识数据408中的每一个可以由网络节点102的各种模块和/或子模块来访问。另外,应当领会的是,在一些实施例中,在网络节点数据402、监视结果数据404、请求监视数据406、和NUMA标识数据408中的每一个中存储的或以其他方式由其表示的数据可以彼此不相互排斥。

[0043] 例如,在一些实施中,在网络节点数据402中存储的数据也可以被存储为监视结果数据404的一部分,和/或反之亦然。照此,尽管由网络节点102利用的各种数据在本文中被描述为具体离散数据,但是在其他实施例中,此类数据可以被组合、聚合和/或以其他方式形成单个或多个数据集合的部分,包括副本拷贝。要进一步领会的是,网络节点102可包括通常在计算装置中找到的另外的和/或备选的组件、子组件、模块、子模块、和/或装置,其为了描述的清晰而未在图4中示出。

[0044] 通信管理模块410(其可以实施为如上所述的硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合)被配置成促进往和来于网络节点102的入站和出站的有线和/或无线网络通信(例如,网络业务、网络分组、网络流等)。为了这样做,通信管理模块410被配置成经由互连结构接收和处理来自其他网络节点102的网络分组。另外,通信管理模块410被配置成经由互连结构准备网络分组和将其传送到其他网络节点102。因此,在一些实施例中,通信管理模块410的至少一部分功能性可以由网络节点102的通信电路220来执行,或者更特定地由通信电路220的HFI 222来执行。在一些实施例中,可用于与结构架构的其他网络节点102通信的数据(诸如IP地址信息、业务信息等)可以被存储在网络节点数据中。

[0045] QoS监视模块420(其可以被实施为如上所述的硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合)被配置成监视网络节点102的各种特性。为了这样做,说明性QoS监视模块420包括资源利用监视模块422、负载平衡监视模块424、和HFI饱和监视模块426。应当领会的是,QoS监视模块420的资源利用监视模块422、负载平衡监视模块424、和HFI饱和监视模块426中的每一个可以单独地实施为硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合。例如,资源利用监视模块422可以实施为硬件组件,而负载平衡监视模块424和/或HFI饱和监视模块426可以实施为虚拟化硬件组件或实施为硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合的某一其他组合。

[0046] 资源利用监视模块422被配置成监视网络节点102的资源(即,物理和/或虚拟组件)的利用等级。在说明性示例中,资源利用监视模块422可以被配置成监视存储器利用等级。为了这样做,在一些实施例中,资源利用监视模块422可以被配置成接收当前由网络节点102的处理器的一个或多个本地高速缓存代理生成的节流信号,其可用于减慢或以其他方式降低对由节流信号指示的给定的存储器类型的注入速率。另外地或备选地,资源利用监视模块422可以被配置成标识资源的当前使用等级以确定被监视资源的饱和等级。

[0047] 负载平衡监视模块424被配置成跨网络节点102的资源(即,物理和/或虚拟组件)监视工作负载的分配。HFI利用监视模块426被配置成监视HFI 222的利用。因此,即使所附连到的资源没有变得饱和,HFI利用监视模块426也可以检测HFI 222的饱和。在说明性示例中,计算节点104之一可以在访问存储节点112的存储装置时使存储节点112之一的HFI 222

饱和。在此类条件下,存储节点112的HFI 222可能变得饱和,而存储节点112的存储装置可能没有被充分利用(即,饱和)。在一些实施例中,监视结果(例如,当前/历史利用值、当前/历史负载平衡信息等)可以存储在监视结果数据404中。

[0048] 节流消息传送模块430(其可以实施为如上所述的硬件,固件,软件,虚拟化硬件,仿真架构和/或其组合)被配置成生成节流消息并将其传送到其他网络节点102。如前所述,某些条件(即,节流条件)可能在网络节点102上存在,使得由网络节点102生成的请求访问网络节点102的本地资源的资源访问请求可能由于其他网络节点102保持对本地被节流的资源的无阻碍的注入速率而变得饥饿。因此,与仅提供本地节流的现有技术不同,网络节点102被配置成检测此类节流条件并生成用于传送到负责或以其他方式促成节流条件的其他网络节点102的节流消息。

[0049] 为了生成节流消息并将其传送到其他网络节点102,说明性节流消息传送模块430包括节流条件检测模块432和传送模式确定模块434。应当了解的是,QoS监视模块420的节流条件检测模块432和传送模式确定模块434中的每个都可以单独地实施为硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合。例如,节流条件检测模块432可以实施为硬件组件,而传送模式确定模块434实施为虚拟化硬件组件或实施为硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合的某一其他组合。

[0050] 节流条件检测模块432被配置成检测是否存在节流条件。为了这样做,节流条件检测模块432可以被配置成将当前服务质量条件(例如,可以由QoS监视模块420确定)与对应阈值相比较。例如,节流条件检测模块432可以被配置成将当前存储器利用等级与存储器饱和阈值相比较。因此,节流条件检测模块432可以在确定当前存储器利用等级超过存储器饱和阈值时检测节流条件。另外地或备选地,节流条件检测模块432可以被配置成处理在网络节点102内部生成的节流信号。如前所述,代理实体(例如,高速缓存代理、归属代理等)可以生成对具体管芯上集群(诸如存储或I/O)的本地节流请求。因此,节流条件检测模块432被配置成解译此类本地节流请求以确定它们是否指示节流条件,由此应当通知其他网络节点102中的一个或多个网络节点采取适当的节流动作。

[0051] 传送模式确定模块434被配置成确定要使用哪个传送模式来传送响应于检测到的节流条件(如被节流条件检测模块432检测到的)而生成的节流消息。为了这样做,传送模式确定模块434被配置成基于负责或以其他方式促成节流条件的标识的其他网络节点102来检测将节流消息传送到网络节点102中的哪一个或多个。例如,传送模式确定模块434可以确定单个网络节点102正在发出太多的存储器访问请求,在这种情况下,传送模式确定模块434可以确定使用单播模式来传送所生成的节流消息。否则,如果传送模式确定模块434确定网络节点102中的多过一个网络节点负责或以其他方式促成当前节流条件,则传送模式确定模块434可以确定使用多播模式来传送所生成的节流消息。

[0052] 如前所述,节流消息传送模块430被配置成响应于节流消息的接收来传送请求另外的网络节点102采取动作的节流消息(例如,节流具体NUMA域的处理器核)。每个网络节点102的每个NUMA域300有对应的NUMA域标识符,其由节流消息传送模块430可用于确定哪个NUMA域300要被节流。因此,网络节点102包括对网络节点102本地的NUMA域300的NUMA域标识符以及其他网络节点102的NUMA域的NUMA域标识符。然而,在一些实施例中,NUMA域标识符可能不是众所周知,诸如在分布式标签目录方案中。在此类实施例中,节流消息传送模块

430可以预测接收网络节点102将对哪个NUMA域300执行响应动作。

[0053] 为了这样做,节流消息传送模块430可以进一步被配置成基于访问NUMA域的应用将在该NUMA域中的某个范围的存储器地址内操作的原理来预测接收网络节点102将对哪个NUMA域300采取行动。网络节点102的管芯上互连接口(例如,图3的点对点接口316之一)被配置成生成对代理实体(诸如高速缓存代理)的请求。因此,管芯上互连接口可以被扩展来使用域预测表来确定哪个NUMA域对应于节流消息,以及NUMA域(例如,NUMA域的处理器核)当前是否遇险(distress)(即,对于该NUMA域已经被节流)。如果NUMA域的组件当前遇险,则可以不向代理实体发出(即,注入)节流消息,直到遇险不再存在并被该代理实体确认为止。因此,域预测表的使用可以允许网络节点102推测另一网络节点102的受影响的NUMA域300。在一些实施例中,域预测表的数据可以被存储在NUMA标识数据408中。

[0054] 域预测表可以包括每个代理实体的标识符、被每个网络节点102已知的NUMA等级,对于每个NUMA域300和/或代理实体访问的最后地址范围(例如,格式化为位掩码)、以及可以每NUMA域300可配置的粒度。在说明性实施例中,粒度对于具体NUMA域300是4GB,并且发送给针对具体NUMA等级的具体代理实体的最后地址(例如,0x78C9657FA)属于地址范围0x700000000 - 0x700000000 + 4GB。如前所述,访问NUMA域的应用会在该NUMA域中的特定范围的存储器地址内进行操作。照此,通过适当地指定粒度,它可以产生更准确的预测,导致高命中率,以及在几个周期内返回结果。因此,在一些实施例中,为了预测针对特定地址和特定代理实体的节流消息的NUMA域,节流消息传送模块430可以被配置成访问域预测表来检索预测的NUMA域作为内容可寻址的存储器(CAM)结构。例如,如果对于应用的用例是要经由公开给NUMA域的存储器来分配存储节点112的10GB存储器块,则如果所选择的粒度是GM,域预测请求很可能会在预测表上命中。

[0055] 在一些实施例中,对于到具体代理实体到具体地址的节流消息的流程可以包括确定最后存储器地址的模数和粒度来预测最后存储器地址属于哪个NUMA域。如果预测的NUMA域请求返回NULL(即,没有NUMA域匹配),则可以假设最接近的NUMA域是NUMA等级0。如前所述,当对于NUMA域的遇险信号是有效的时,处理器核不向代理实体发送事务,并且只在遇险信号是无效的和被代理实体确认之后才发出事务。因此,如果遇险信号对于预测的NUMA域的结果是有效的,则处理器核不向代理实体发送事务,直到遇险信号是无效的和被代理实体确认为止。另外,在一些实施例中,代理实体可以执行系统地址解码,更新适当的计数器(例如,节流请求计数器),并在必要时生成遇险信号。此外,根据依据代理实体已经返回确认而接收到的反馈和对于该特定域预测请求的NUMA域标识符来更新预测表。

[0056] 应当领会的是,本文描述的基于QoS的节流方案的目标光纤架构是针对具有数百个网络节点102的规模的企业系统。因此,在具有更大规模的此类实施例(诸如数千节点的高性能计算(HPC)的规模的实施例)中,由于可以在其中传送的大量消息,多播模式对于实现可以不是理想的。然而,结构架构的网络节点102的子域(例如,仅由与特定网络交换机116连接的那些网络节点102组成)可以诸如通过使用特定的多播拓扑来定义,以便于将节流消息仅传播到网络节点102的子集。应该进一步领会的是,多播模式可以是不可靠的多播。如前所述,节流消息在节流条件存在的持续时间内周期性地被传送,从而取消对节流消息的接收进行确认的需要。可靠性可以诸如通过在流程中增加接收确认被改进;然而,这种可靠性改进可能会在结构中增加更多压力。

[0057] 在一些实施例中,节流消息传送模块430(例如,节流条件检测模块432和/或传送模式确定模块434)可以利用请求监视表来确定什么时候生成节流消息和/或生成的节流消息要被传送到哪些网络节点102。在说明性示例中,节流消息传送模块430可以被配置成考虑针对具体网络节点102的本地NUMA域300的外部事务。如前所述,每个NUMA域300具有对应的NUMA域标识符,其由节流消息传送模块430可用于确定接收到的节流消息对应哪个NUMA域300、以及随每次访问而递增的请求计数器。

[0058] 在一些实施例中,NUMA域标识符、请求计数器的值以及其他值(例如,节流消息请求类型的枚举值)可以被存储在模型特定寄存器(MSR)中。因此,节流消息传送模块430可以被配置成读取请求计数器的值来确定请求计数器是否超过阈值。应当领会的是,MSR值可以在操作或引导时间(例如,使用环零功能)期间被配置,并且可以公开给网络节点102的操作系统。

[0059] 在一些实施例中,请求计数器可以被存储在请求监视表中,其包括从其接收节流消息的网络节点102的标识符、请求计数器的当前值、NUMA域标识符、和阈值。在一些实施例中,请求监视表的数据可以被存储在请求监视数据406中。如果请求计数器超过阈值,则节流消息传送模块430可以被配置成生成用于以单播模式传送的节流消息(即,到只负责请求计数器的当前状态的网络节点102)。另外,节流消息传送模块430可以被配置成在内部接收到节流信号(诸如从缓存代理)时,生成用于以多播模式传送的节流消息(即,到向具体NUMA域发出事务的所有其他网络节点102)。如前所述,节流消息传送模块430被配置成在检测到节流条件时周期性地生成节流消息。

[0060] 应当领会的是,诸如NUMA域标识符、不同网络节点102中的MSR等的系统配置应该整体地进行以确保一致性。因此,系统配置应该在系统引导时间时(例如,当执行路由系统地址解码方案时)被强制执行来确保在节流消息中传达的信息跨不同网络节点102是一致的。例如,在其中计算节点(1)106的NUMA域标识符正在被传播到计算节点(2)108的一实施例中,计算节点(2)108应该已经知道哪个NUMA域标识符对应于计算节点(1)106的具体NUMA域300。在一些实施例中,其他网络节点102的NUMA域标识符可以被存储在NUMA标识数据408中。

[0061] 节流消息接收模块440(其可以实施为如上所述的硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合)被配置成接收和处理来自其他网络节点102的节流消息。为了这样做,说明性节流消息接收模块440包括节流类型标识模块442和NUMA目标标识模块444。应当领会的是,节流消息接收模块440的节流类型标识模块442和NUMA目标标识模块444中的每一个可以单独地实施为硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合。例如,节流类型标识模块442可以实施为硬件组件,而NUMA目标标识模块444实施为虚拟化硬件组件或实施为硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合的某一其他组合。

[0062] 节流类型标识模块442被配置成标识与接收到的节流消息相关联的类型。如前所述,与节流消息相关联的节流消息请求类型可以包括存储器节流请求、I/O节流请求、加速器节流处理请求、HFI饱和节流请求等。在一些实施例中,节流消息请求类型可以是枚举型,使得它们可以映射到具体动作。另外,一些实施例,节流消息请求类型的枚举值可以被存储在将枚举值映射到对应动作的节流动作表中。NUMA目标标识模块444被配置成标识与接收到的节流消息相关联的NUMA域目标或其组件。

[0063] 节流响应执行模块450(其可以实施为如上所述的硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合)被配置成响应与已经从另一网络节点102接收到节流消息而采取动作。为了这样做,说明性节流响应执行模块450包括处理器核节流执行模块452、软件中断执行模块454和HFI节流执行模块456。应当领会的是,节流响应执行模块450的处理器核节流执行模块452、软件中断执行模块454、和HFI节流执行模块456中的每一个可以单独地实施为硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合。例如,处理器核节流执行模块452可以实施为硬件组件,而软件中断执行模块454和/或HFI节流执行模块456可以实施为虚拟化硬件组件或实施为硬件、固件、软件、虚拟化硬件、仿真架构和/或其组合的某一其他组合。

[0064] 处理器核节流执行模块452被配置成响应于接收传播的节流消息而节流处理器核。为了这样做,处理器核节流执行模块452被配置成将接收到的节流消息转换为由网络节点102架构支持的对应的管芯上互连节流信号以降低外部传送的访问请求的注入速率。软件中断执行模块454被配置成响应于已经接收到的软件中断请求节流消息而执行软件中断。为了这样做,在此类实施例中,节流消息经由软件中断被传播到软件堆,其中软件堆支持负载平衡和注入控制机制。

[0065] HFI节流执行模块456被配置成基于接收到的节流消息的类型来节流在HFI 222处的注入。换句话说,HFI 222负责降低注入速率或完全停止注入。因此,此类响应可以是对于由网络节点102的结构架构不支持的节流消息类型的合适解决方案。应当领会的是,网络节点102的处理器核和其他注入器未被节流。

[0066] 现在参考图5,在使用中,网络节点102(例如,图1的网络节点102之一)可以执行用于处理来自远程网络节点(即,结构架构的网络节点102的另一个)的本地存储器请求的方法500。方法500开始于框502,其中网络节点102确定是否已经从远程网络节点接收到存储器访问请求。如果没有,则方法500循环回到框502以确定是否已经从远程网络节点接收到存储器访问请求;否则,方法500前进到框504。在框504中,网络节点102将接收到的远程存储器访问请求插入到网络节点102的共享缓冲器中。应当领会的是,在某些条件下,共享缓冲器可能在网络节点102可以将接收到的远程存储器访问请求插入到共享缓冲器之前的一段时间内是满的。

[0067] 在框506中,网络节点102确定是否处理接收到的请求(例如,从共享缓冲器弹出相应的条目并处理该请求)。如果是,则方法500前进到框508,其中网络节点102执行响应于接收到的远程存储器访问请求的动作。例如,在框510中,网络节点102响应于远程存储器访问请求已经请求将数据存储到网络节点102的存储器(例如,主存储器216)中而传送请求数据。备选地,在框512中,网络节点102可以存储与远程存储器访问请求一起接收的数据。在一些实施例中,在框514中,网络节点102响应已经接收/处理的远程存储器访问请求而传送确认。

[0068] 现在参考图6,在使用中,网络节点102(例如,图1的网络节点102之一)可以执行用于访问远程网络节点(即,结构架构的网络节点102中的另一个)的存储器的方法600。方法600开始于框602,其中网络节点102确定是否访问位于另一网络节点102中的存储器。例如,网络节点102可以正检索在远程存储器中复制的数据(即,远程网络节点的存储器),执行在一个或多个远程网络节点上利用分布式数据结构的应用,采用日志运送(即,依赖存储在远

程网络节点上的用于故障恢复的日志或微日志),或执行请求访问远程网络节点的存储器的一些其他操作。

[0069] 如果不是,则方法600循环回到框602来再次确定是否访问位于另一网络节点102中的存储器;否则,方法600前进到框604。在框604中,网络节点102生成远程存储器访问请求,其包括可用于检索或存储远程存储器访问请求的数据的存储器地址信息。另外,在框606中,网络节点102包括网络节点102的源标识信息。在框608中,网络节点102将存储器访问请求插入到消息传送队列中。

[0070] 在框610中,网络节点102确定对应于远程存储器访问请求正从其请求访问的组件的注入速率是否由于从远程网络节点接收的节流消息(参见例如,图7的方法700,其针对生成用于到一个或多个远程网络节点的外部传送的节流消息)而被节流。如果没有,则方法600分支到框612,其中网络节点102以未节流的注入速率传送远程存储器访问请求;否则,方法600分支到框614,其中网络节点102以节流速率的传送远程存储器访问请求。

[0071] 现在参考图7,在使用中,网络节点102(例如,图1的网络节点102之一)可以执行方法700,其用于生成用于到一个或多个远程网络节点(即,其他远程网络节点的一个或多个)的外部传送的节流消息。方法700开始于框702,其中网络节点102监视网络节点102的服务质量等级。例如,在框704中,在一些实施例中,网络节点102监视网络节点102的资源(例如,存储器、处理器、NUMA域的组件等)的利用等级。另外地或备选地,在框706中,在一些实施例中,网络节点102监视跨网络节点102的组件分布的工作负载的分布。在框708中,在一些实施例中,网络节点102另外或备选地监视网络节点102的HFI 222的饱和等级。如前所述,在一些实施例中,网络节点102可以依赖于请求监视表来确定什么时候要为具体NUMA域300生成节流消息。

[0072] 也如前所述,网络节点102上可以存在某些条件(即,节流条件),使得由网络节点102生成的请求对网络节点102的本地资源进行访问的资源访问请求可能由于远程网络节点102保持对被本地节流的资源的无阻碍注入速率而变得饥饿。因此,在框710中,网络节点102确定是否由于在框702中执行的监视服务质量而已检测到(即,当前存在)节流条件(例如,网络节点102的组件的拥塞、饱和、过度利用、工作负载分配不公等)。

[0073] 如果网络节点102确定不存在节流条件,则方法700循环回到框702以继续监视网络节点102的服务质量等级;否则,方法700前进到框712,其中网络节点102生成节流消息。在框714中,网络节点102包括具有节流消息的节流消息请求类型指示符。如前所述,与节流消息相关联的节流消息请求类型可以包括存储器节流请求、I/O节流请求、加速器节流处理请求、HFI饱和节流请求等。另外,在框716中,网络节点102包括节流消息源指示符。节流消息源指示符可以包括对于其已经检测到节流条件的组件的标识符(例如,NUMA域标识符、HFI标识符)和/或网络节点102的标识符。

[0074] 在框718中,网络节点102标识要接收在框712中生成的节流消息的一个或多个目标网络节点(即,结构架构的其他网络节点102中的一个或多个)。如前所述,在一些实施例中,网络节点102可以依赖于请求监视表来确定所述一个或多个目标网络节点。在框720中,网络节点102将生成的节流消息传送到在框718中标识的所述一个或多个目标网络节点。为了这样做,在框722中,网络节点102基于对应于每个目标网络节点的循环速率来传送所生成的节流消息。

[0075] 取决于在框718中标识的目标网络节点的数量,网络节点102可以在框724中经由多播传输(即,多于一个目标网络节点)或者在框726中经由单播传输(即,单个目标网络节点)传送生成的节流消息。另外,在一些实施例中,在框728中,网络节点102可以经由OSI模型的传输层传送生成的节流消息。为了这样做,在一些实施例中,可以使用促进节流消息的转移的新类型的虚拟信道来扩展结构,以便于将节流消息从结构的现有的信道隔离。此类实施例可以经由在结构内选择最快路径的新类型的物理线路来实现,以便于尽可能快地传递节流消息。

[0076] 如前所述,在检测到的节流条件的过程中周期性地传送节流消息。照此,方法700可以监视相对于该特定节流条件的服务质量等级,并且由于该特定服务质量等级监视而迭代方法700。

[0077] 现在参考图8,在使用中,网络节点102(例如,图1的网络节点102之一)可以执行用于处理从远程网络节点(即,其他网络节点102之一)接收到的节流消息的方法800。方法800开始于框802,其中网络节点102确定是否已从远程网络节点接收到节流消息。如果没有,则方法800循环回到框802以再次确定是否已从远程网络节点接收到节流消息;否则,方法800前进到框804,其中网络节点102标识与在框802中接收到的节流消息相关联的信息。

[0078] 例如,在框806中,网络节点102标识节流消息的类型。如前所述,与节流消息相关联的节流消息请求类型可以包括存储器节流请求、I/O节流请求、加速器节流处理请求、HFI饱和节流请求等。另外,在框808中,网络节点102标识节流消息的源。节流消息的源可以包括标识从其接收到的节流消息的目标网络节点的信息。另外,节流消息的源可以包括标识从其接收了节流任务的远程网络节点的组件的组件标识符(例如,NUMA标识符)。在一些实施例中,节流消息可以另外包括接收网络节点102的组件信息,其可用于标识哪些网络节点102资源(远程存储器访问正从其被请求)要被节流。

[0079] 在框810中,网络节点102基于接收到的节流消息来执行动作,诸如可以基于在框806中标识的节流消息的类型来执行动作。例如,在框812中,网络节点102可以由网络节点102的HFI 222通过自我节流请求来降低对于正被传送到(即,针对)远程网络节点的共享资源访问请求的注入速率。在另一个示例中,在框814中,网络节点102可以通过使用现有的节流方案来节流网络节点102处理器核。为了这样做,网络节点102可以经由对应的管芯上互连(例如,处理器202的管芯上互连306、处理器208的管芯上互连312等)将接收到的节流消息传播到代理实体(例如,高速缓存代理)以通过使用现有的节流方案来节流网络节点102的处理器核。在又一示例中,在框816中,在其中软件堆支持负载平衡和注入控制机制的此类实施例中,网络节点102可以经由软件中断将接收到的节流消息传播到软件堆。

[0080] 应当领会的是,方法500、600、700和800中的一个或多个方法的至少一部分可以通过网络节点102的HFI 222执行。应当进一步领会的是,在一些实施例中,方法500、600、700和800中的一个或多个方法可以实施为存储在计算机可读介质上的各种指令,其可以由处理器(例如,处理器202、处理器208等)、HFI 222、和/或网络节点102的其他组件来执行以促使网络节点102执行方法500、600、700和800。计算机可读介质可以实施为能够被网络节点102读取的任何类型的介质,其包括但不限于主存储器216、数据存储装置218、HFI 222的安全存储器(未示出)、网络节点102的其他存储器或数据存储装置、由网络节点102的外围装置可读取的便携式介质、和/或其他介质。

[0081] 示例

[0082] 以下提供本文公开的技术的说明性示例。这些技术的一实施例可以包括下面描述的示例中的任何一个或多个、以及任何组合。

[0083] 示例1包括一种用于在结构架构中基于服务质量进行节流的网络节点,其中所述网络节点是所述结构架构的多个互连网络节点之一,所述网络节点包括处理器;主机结构接口(HFI),所述HFI用于促进通过所述结构架构的互连结构在所述多个互连网络节点之间数据传送;和一个或多个数据存储装置,所述一个或多个数据存储装置中存储了多个指令,当由所述处理器执行时,所述多个指令促使所述网络节点监视所述网络节点的服务质量等级;基于监视的服务质量等级的结果来检测节流条件;响应于已检测到所述节流条件,基于与检测到的所述节流条件相关联的请求类型来生成节流消息;以及将生成的节流消息传送到经由所述互连结构通信地耦合到所述网络节点的所述多个互连网络节点中的一个或多个。

[0084] 示例2包括示例1的主题,并且其中监视所述网络节点的服务质量等级包括监视所述网络节点的一个或多个资源的利用等级。

[0085] 示例3包括示例1和2中的任一项的主题,并且其中所述网络节点的所述一个或多个资源包括所述处理器、所述一个或多个数据存储装置、或所述HFI中的至少一项。

[0086] 示例4包括示例1-3中的任一项的主题,并且其中所述网络节点的所述一个或多个资源包括多个非统一存储器访问(NUMA)域,其中所述多个NUMA域中的每一个包括所述处理器的处理器核的分配的部分和所述一个或多个数据存储装置的分配的部分,并且其中所述多个NUMA域中的每一个经由所述处理器的管芯上互连通信地耦合到所述HFI。

[0087] 示例5包括示例1-4中的任一项的主题,并且其中监视所述网络节点的所述服务质量等级包括监视工作负载分配。

[0088] 示例6包括示例1-5中的任一项的主题,并且其中监视所述网络节点的所述服务质量等级包括监视所述HFI的饱和等级。

[0089] 示例7包括示例1-6中的任一项的主题,并且其中监视所述网络节点的所述服务质量等级包括监视从所述处理器的代理实体接收到的节流信号。

[0090] 示例8包括示例1-7中的任一项的主题,并且其中所述多个互连网络节点包括一个或多个计算节点和一个或多个存储节点。

[0091] 示例9包括示例1-8中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括经由单播传送将生成的节流消息传送到所述多个互连网络节点中的一个。

[0092] 示例10包括示例1-9中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括经由多播传送将生成的节流消息传送到所述多个互连网络节点中的多于一个。

[0093] 示例11包括示例1-10中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括经由开放系统互连(OSI)模型的传输层将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个。

[0094] 示例12包括示例1-11中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括在检测到的节流条件的持续时间内以周期性

注入速率将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个。

[0095] 示例13包括一种用于在结构架构中基于服务质量进行节流的网络节点,其中所述网络节点是所述结构架构的多个互连网络节点之一,所述网络节点包括服务质量监视电路,所述服务质量监视电路用于在所述结构架构的互连结构上监视所述网络节点的主机结构接口(HFI)与其他互连网络节点的一个或多个HFI之间的所述网络节点的服务质量等级;节流消息传输电路,所述节流消息传输电路用于(i)基于监视的服务质量等级的结果来检测节流条件,(ii)响应于已检测到所述节流条件,基于与检测到的所述节流条件相关联的请求类型来生成节流消息,以及(iii)将生成的节流消息传送到经由所述互连结构通信地耦合到所述网络节点的所述多个互连网络节点中的一个或多个。

[0096] 示例14包括示例13的主题,并且其中监视所述网络节点的服务质量等级包括监视所述网络节点的一个或多个资源的利用等级。

[0097] 示例15包括示例13和14中的任一项的主题,并且其中所述网络节点的所述一个或多个资源包括所述网络节点的处理器、所述网络节点的一个或多个数据存储装置、或所述HFI中的至少一项。

[0098] 示例16包括示例13-15中的任一项的主题,并且其中所述网络节点的所述一个或多个资源包括多个非统一存储器访问(NUMA)域,其中所述多个NUMA域中的每一个包括所述处理器的处理器核的分配的部分和所述一个或多个数据存储装置的分配的部分,并且其中所述多个NUMA域中的每一个经由所述处理器的管芯上互连通信地耦合到所述HFI。

[0099] 示例17包括示例13-16中的任一项的主题,并且其中监视所述网络节点的所述服务质量等级包括监视工作负载分配。

[0100] 示例18包括示例13-17中的任一项的主题,并且其中监视所述网络节点的所述服务质量等级包括监视所述HFI的饱和等级。

[0101] 示例19包括示例13-18中的任一项的主题,并且其中监视所述网络节点的所述服务质量等级包括监视从所述处理器的代理实体接收到的节流信号。

[0102] 示例20包括示例13-19中的任一项的主题,并且其中所述多个互连网络节点包括一个或多个计算节点和一个或多个存储节点。

[0103] 示例21包括示例13-20中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括经由单播传送将生成的节流消息传送到所述多个互连网络节点中的一个。

[0104] 示例22包括示例13-21中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括经由多播传送将生成的节流消息传送到所述多个互连网络节点中的多于一个。

[0105] 示例23包括示例13-22中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括经由开放系统互连(OSI)模型的传输层将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个。

[0106] 示例24包括示例13-23中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括在检测到的节流条件的持续时间内以周期性注入速率将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个。

[0107] 示例25包括一种用于跨结构架构基于服务质量进行节流的方法,所述方法包括通

过所述结构架构的网络节点的主机结构接口(HFI)监视所述网络节点的服务质量等级,其中所述网络节点是所述结构架构的多个互连网络节点之一,其中所述多个互连网络节点中的每一个通过所述结构架构的互连结构被互连;通过所述HFI基于监视的服务质量等级的结果检测节流条件;响应于已检测到所述节流条件,通过所述HFI基于与检测到的所述节流条件相关联的请求类型生成节流消息;以及通过所述HFI将生成的节流消息传送到经由所述互连结构通信地耦合到所述网络节点的所述多个互连网络节点中的一个或多个。

[0108] 示例26包括示例25的主题,并且其中监视所述网络节点的所述服务质量等级包括监视所述网络节点的一个或多个资源的利用等级。

[0109] 示例27包括示例25和26中的任一项的主题,并且其中所述网络节点的一个或多个资源的利用等级包括监视所述网络节点的处理器、所述网络节点的一个或多个数据存储装置、或所述HFI中的至少一项。

[0110] 示例28包括示例25-27中的任一项的主题,并且其中监视所述网络节点的一个或多个资源的所述利用等级包括监视多个非统一存储器访问(NUMA)域中的一个或多个,其中所述多个NUMA域中的每一个包括所述网络节点的处理器的处理器核的分配的部分和所述网络节点的一个或多个数据存储装置的分配的部分,并且其中所述多个NUMA域中的每一个经由所述处理器的管芯上互连通信地耦合到所述HFI。

[0111] 示例29包括示例25-28中的任一项的主题,并且其中监视所述网络节点的所述服务质量等级包括监视工作负载分配。

[0112] 示例30包括示例25-29中的任一项的主题,并且其中监视所述网络节点的所述服务质量等级包括监视所述HFI的饱和等级。

[0113] 示例31包括示例25-30中的任一项的主题,并且其中监视所述网络节点的所述服务质量等级包括监视从所述处理器的代理实体接收到的节流信号。

[0114] 示例32包括示例25-31中的任一项的主题,并且其中所述多个互连网络节点包括一个或多个计算节点和一个或多个存储节点。

[0115] 示例33包括示例25-32中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括经由单播传送将生成的节流消息传送到所述多个互连网络节点中的一个。

[0116] 示例34包括示例25-33中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括经由多播传送将生成的节流消息传送到所述多个互连网络节点中的多于一个。

[0117] 示例35包括示例25-34中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括经由开放系统互连(OSI)模型的传输层将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个。

[0118] 示例36包括示例25-35中的任一项的主题,并且其中将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个包括在检测到的节流条件的持续时间内以周期性注入速率将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个。

[0119] 示例37包括一种网络节点,所述网络节点包括处理器;以及存储器,所述存储器上存储有多个指令,所述多个指令当由所述处理器执行时促使所述网络节点执行示例25-36中的任一项的方法。

[0120] 示例38包括一个或多个机器可读存储介质,包括存储在所述一个或多个机器可读存储介质上的多个指令,多个指令响应于被执行而导致网络节点执行示例25-36中的任一项的方法。

[0121] 示例39包括一种用于在结构架构中基于服务质量进行节流的网络节点,其中所述网络节点是所述结构架构的多个互连网络节点之一,所述网络节点包括用于在所述网络节点的主机结构接口(HFI)监视所述网络节点的服务质量等级的部件,其中所述网络节点是所述结构架构的多个互连网络节点之一,其中所述多个互连网络节点中的每一个通过所述结构架构的互连结构被互连;用于基于监视的服务质量等级的结果检测节流条件的部件;用于响应于已检测到所述节流条件而基于与检测到的所述节流条件相关联的请求类型生成节流消息的部件;以及用于将生成的节流消息传送到经由所述互连结构通信地耦合到所述网络节点的多个互连网络节点中的一个或多个的部件。

[0122] 示例40包括示例39的主题,并且其中用于监视所述网络节点的所述服务质量等级的部件包括用于监视所述网络节点的一个或多个资源的利用等级的部件。

[0123] 示例41包括示例39和40中的任一项的主题,并且用于其中所述网络节点的一个或多个资源的利用等级的部件包括用于监视所述网络节点的处理器、所述网络节点的一个或多个数据存储装置、或所述HFI中的至少一项的部件。

[0124] 示例42包括示例39-41中的任一项的主题,并且其中用于监视所述网络节点的一个或多个资源的所述利用等级的部件包括用于监视多个非统一存储器访问(NUMA)域中的一个或多个的部件,其中所述多个NUMA域中的每一个包括所述网络节点的处理器的处理器核的分配的部分和所述网络节点的一个或多个数据存储装置的分配的部分,并且其中所述多个NUMA域中的每一个经由所述处理器的管芯上互连通信地耦合到所述HFI。

[0125] 示例43包括示例39-42中的任一项的主题,并且其中用于监视所述网络节点的所述服务质量等级的部件包括用于监视工作负载分配的部件。

[0126] 示例44包括示例39-43中的任一项的主题,并且其中用于监视所述网络节点的所述服务质量等级的部件包括用于监视所述HFI的饱和等级的部件。

[0127] 示例45包括示例39-44中的任一项的主题,并且其中用于监视所述网络节点的所述服务质量等级的部件包括用于监视从所述处理器的代理实体接收到的节流信号的部件。

[0128] 示例46包括示例39-45中的任一项的主题,并且其中所述多个互连网络节点包括一个或多个计算节点和一个或多个存储节点。

[0129] 示例47包括示例39-46中的任一项的主题,并且其中用于将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个的部件包括用于经由单播传送将生成的节流消息传送到所述多个互连网络节点中的一个的部件。

[0130] 示例48包括示例39-47中的任一项的主题,并且其中用于将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个的部件包括用于经由多播传送将生成的节流消息传送到所述多个互连网络节点中的多于一个的部件。

[0131] 示例49包括示例39-48中的任一项的主题,并且其中用于将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个的部件包括用于经由开放系统互连(OSI)模型的传输层将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个的部件。

[0132] 示例50包括示例39-49中的任一项的主题,并且其中用于将生成的节流消息传送

到所述多个互连网络节点中的所述一个或多个的部件包括用于在检测到的节流条件的持续时间内以周期性注入速率将生成的节流消息传送到所述多个互连网络节点中的所述一个或多个的部件。

[0133] 示例51包括一种用于在结构架构中基于服务质量进行节流的网络节点,其中所述网络节点是所述结构架构的多个互连网络节点之一,所述网络节点包括处理器;用于促进在所述多个互连网络节点之间的数据传送的主机结构接口(HFI);以及一个或多个数据存储装置,所述一个或多个数据存储装置中存储了多个指令,所述多个指令当由所述处理器执行时,促使所述网络节点传送访问请求,所述访问请求用于对经由所述结构架构的互连结构通信地耦合到所述网络节点的所述多个互连网络节点之一的共享资源的访问;从所述多个互连网络节点之一接收节流消息;标识与接收到的节流消息相关联的信息;并且基于所标识的信息对所述网络节点的一个或多个资源执行节流动作。

[0134] 示例52包括示例51的主题,并且其中接收所述节流消息包括经由开放系统互连(OSI)模型的传输层接收所述节流消息。

[0135] 示例53包括示例51和52中的任一项的主题,并且其中标识与接收到的节流消息相关联的所述信息包括标识接收到的节流消息的请求类型和接收到的节流消息的源中的至少一项。

[0136] 示例54包括示例51-53中的任一项的主题,并且其中接收到的节流消息的所述请求类型包括存储器节流请求、I/O节流请求、加速器节流处理请求、或HFI饱和节流请求中的一个。

[0137] 示例55包括示例51-54中的任一项的主题,并且其中执行所述节流动作包括降低针对所述多个互连网络节点之一的共享资源访问请求的注入速率。

[0138] 示例56包括示例51-55中的任一项的主题,并且其中执行所述节流动作包括节流所述网络节点的所述处理器的处理器核。

[0139] 示例57包括示例51-56中的任一项的主题,并且其中节流所述网络节点的所述处理器的所述处理器核包括将接收到的节流消息传播到所述处理器的代理实体,以用于转换成现有的节流信号。

[0140] 示例58包括示例51-57中的任一项的主题,并且其中执行所述节流动作包括经由软件中断将接收到的节流消息传播到软件堆。

[0141] 示例59包括示例51-58中的任一项的主题,并且其中所述网络节点的所述一个或多个资源包括所述处理器、所述一个或多个数据存储装置或所述HFI中的至少一项。

[0142] 示例60包括示例51-59中的任一项的主题,并且其中所述网络节点的所述一个或多个资源包括多个非统一存储器访问(NUMA)域,其中所述多个NUMA域中的每一个包括所述处理器的处理器核的分配的部分和所述一个或多个数据存储装置的分配的部分,并且其中所述多个NUMA域中的每一个经由所述处理器的管芯上互连通信地耦合到所述HFI。

[0143] 示例61包括一种用于在结构架构中基于服务质量进行节流的网络节点,其中所述网络节点是所述结构架构的多个互连网络节点之一,所述网络节点包括通信管理电路,所述通信管理电路用于传送访问请求,所述访问请求用于对经由所述结构架构的互连结构通信地耦合到所述网络节点的所述多个互连网络节点之一的共享资源的访问;节流消息接收电路,所述节流消息接收电路用于(i)从所述多个互连网络节点之一的主机结构接口(HFI)

接收节流消息;以及(ii)标识与接收到的节流消息相关联的信息;以及节流相应执行电路,所述节流相应执行电路用于基于所标识的信息对所述网络节点的一个或多个资源执行节流动作。

[0144] 示例62包括示例61的主题,并且其中接收所述节流消息包括经由开放系统互连(OSI)模型的传输层接收所述节流消息。

[0145] 示例63包括示例61和62中的任一项的主题,并且其中标识与接收到的节流消息相关联的所述信息包括标识接收到的节流消息的请求类型和接收到的节流消息的源中的至少一项。

[0146] 示例64包括示例61-63中的任一项的主题,并且其中接收到的节流消息的所述请求类型包括存储器节流请求、I/O节流请求、加速器节流处理请求、或HFI饱和节流请求中的一个。

[0147] 示例65包括示例61-64中的任一项的主题,并且其中执行所述节流动作包括降低针对所述多个互连网络节点之一的共享资源访问请求的注入速率。

[0148] 示例66包括示例61-65中的任一项的主题,并且其中执行所述节流动作包括节流所述网络节点的所述处理器的处理器核。

[0149] 示例67包括示例61-66中的任一项的主题,并且其中节流所述网络节点的所述处理器的所述处理器核包括将接收到的节流消息传播到所述处理器的代理实体,以用于转换成现有的节流信号。

[0150] 示例68包括示例61-67中的任一项的主题,并且其中执行所述节流动作包括经由软件中断将接收到的节流消息传播到软件堆。

[0151] 示例69包括示例61-68中的任一项的主题,并且其中所述网络节点的所述一个或多个资源包括所述处理器、所述一个或多个数据存储装置或所述HFI中的至少一项。

[0152] 示例70包括示例61-69中的任一项的主题,并且其中所述网络节点的所述一个或多个资源包括多个非统一存储器访问(NUMA)域,其中所述多个NUMA域中的每一个包括所述处理器的处理器核的分配的部分和所述一个或多个数据存储装置的分配的部分,并且其中所述多个NUMA域中的每一个经由所述处理器的管芯上互连通信地耦合到所述HFI。

[0153] 示例71包括一种用于跨结构架构基于服务质量进行节流的方法,所述网络节点包括通过网络节点的主机结构接口(HFI)传送访问请求,所述访问请求用于对经由所述结构架构的互连结构通信地耦合到所述网络节点的所述多个互连网络节点之一的共享资源的访问;通过所述HFI接收来自所述多个互连网络节点之一的节流消息;通过所述HFI标识与接收到的节流消息相关联的信息;通过所述HFI基于所标识的信息对所述网络节点的一个或多个资源执行节流动作。

[0154] 示例72包括示例71的主题,并且其中接收所述节流消息包括经由开放系统互连(OSI)模型的传输层接收所述节流消息。

[0155] 示例73包括示例71和72中的任一项的主题,并且其中标识与接收到的节流消息相关联的所述信息包括标识接收到的节流消息的请求类型和接收到的节流消息的源中的至少一项。

[0156] 示例74包括示例71-73中的任一项的主题,并且其中标识接收到的节流消息的所述请求类型包括标识存储器节流请求、I/O节流请求、加速器节流处理请求、或HFI饱和节流

请求中的一个。

[0157] 示例75包括示例71-74中的任一项的主题,并且其中执行所述节流动作包括降低针对所述多个互连网络节点之一的共享资源访问请求的注入速率。

[0158] 示例76包括示例71-75中的任一项的主题,并且其中执行所述节流动作包括节流所述网络节点的处理器的处理器核。

[0159] 示例77包括示例71-76中的任一项的主题,并且其中节流所述网络节点的所述处理器的所述处理器核包括将接收到的节流消息传播到所述处理器的代理实体,以用于转换成现有的节流信号。

[0160] 示例78包括示例71-77中的任一项的主题,并且其中执行所述节流动作包括经由软件中断将接收到的节流消息传播到软件堆。

[0161] 示例79包括示例71-78中的任一项的主题,并且其中对所述网络节点的所述一个或多个资源执行所述节流动作包括对所述网络节点的处理器、所述网络节点的一个或多个数据存储装置或者所述HFI中的至少一项执行所述节流动作。

[0162] 示例80包括示例71-79中的任一项的主题,并且其中对所述网络节点的所述一个或多个资源执行所述节流动作包括对多个非统一存储器访问(NUMA)域中的至少一个执行所述节流动作,其中所述多个NUMA域中的每一个包括所述网络节点的处理器的处理器核的分配的部分和所述网络节点的一个或多个数据存储装置的分配的部分,并且其中所述多个NUMA域中的每一个经由所述处理器的管芯上互连通信地耦合到所述HFI。

[0163] 示例81包括一种网络节点,所述网络节点包括处理器;以及存储器,所述存储器上存储有多个指令,所述多个指令当由所述处理器执行时促使所述网络节点执行示例71-80中的任一项的方法。

[0164] 示例82包括一个或多个机器可读存储介质,包括存储在所述一个或多个机器可读存储介质上的多个指令,所述多个指令响应于被执行而导致网络节点执行示例71-80中的任一项的方法。

[0165] 示例83包括一种用于在结构架构中基于服务质量进行节流的网络节点,其中所述网络节点是所述结构架构的多个互连网络节点之一,所述网络节点包括用于通过所述网络节点的主机结构接口(HFI)传送访问请求的部件,所述访问请求用于对经由所述结构架构的互连结构通信地耦合到所述网络节点的所述多个互连网络节点之一的共享资源的访问;用于通过所述HFI接收来自所述多个互连网络节点之一的节流消息的部件;用于通过HFI标识与接收到的节流消息相关联的信息的部件;用于通过HFI基于所标识的信息对所述网络节点的一个或多个资源执行节流动作的部件。

[0166] 示例84包括示例83的主题,并且其中用于接收所述节流消息包括经由开放系统互连(OSI)模型的传输层接收所述节流消息的部件。

[0167] 示例85包括示例83和84中的任一项的主题,并且其中用于标识与接收到的节流消息相关联的所述信息的部件包括用于标识接收到的节流消息的请求类型和接收到的节流消息的源中的至少一项的部件。

[0168] 示例86包括示例83-85中的任一项的主题,并且其中用于标识接收到的节流消息的所述请求类型的部件包括用于标识存储器节流请求、I/O节流请求、加速器节流处理请求、或HFI饱和节流请求中的一个的部件。

[0169] 示例87包括示例83-86中的任一项的主题,并且其中用于执行所述节流动作的部件包括用于降低针对所述多个互连网络节点之一的共享资源访问请求的注入速率的部件。

[0170] 示例88包括示例83-87中的任一项的主题,并且其中用于执行所述节流动作的部件包括用于节流所述网络节点的处理器的处理器核的部件。

[0171] 示例89包括示例83-88中的任一项的主题,并且其中用于节流所述网络节点的所述处理器的所述处理器核的部件包括用于将接收到的节流消息传播到所述处理器的代理实体以用于转换成现有的节流信号的部件。

[0172] 示例90包括示例83-89中的任一项的主题,并且其中用于执行所述节流动作的部件包括用于经由软件中断将接收到的节流消息传播到软件堆的部件。

[0173] 示例91包括示例83-90中的任一项的主题,并且其中用于对所述网络节点的所述一个或多个资源执行所述节流动作的部件包括用于对所述网络节点的处理器、所述网络节点的一个或多个数据存储装置或者所述HFI中的至少一项执行所述节流动作的部件。

[0174] 示例92包括示例83-91中的任一项的主题,并且其中用于对所述网络节点的所述一个或多个资源执行所述节流动作的部件包括用于对多个非统一存储器访问(NUMA)域中的至少一个执行所述节流动作的部件,其中所述多个NUMA域中的每一个包括所述网络节点的处理器的处理器核的分配的部分和所述网络节点的一个或多个数据存储装置的分配的部分,并且其中所述多个NUMA域中的每一个经由所述处理器的管芯上互连通信地耦合到所述HFI。

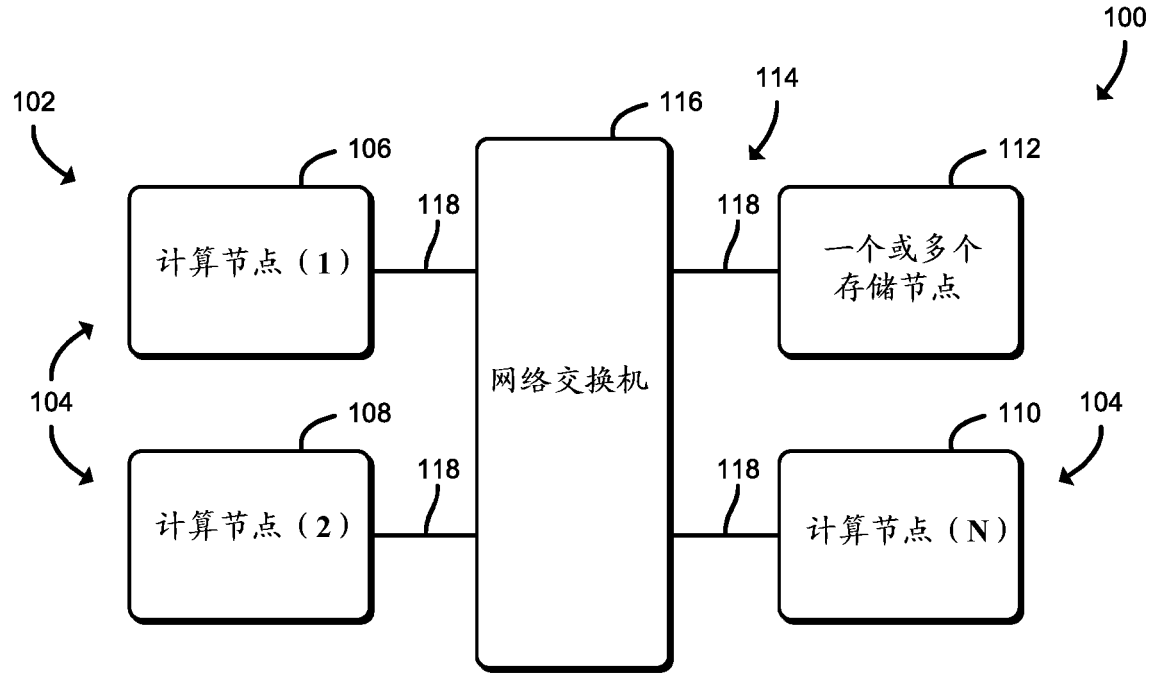

100
102
106
116
114
112
118
118
计算节点（1）
网络交换机
一个或多个
存储节点
104
108
110
104
118
118
计算节点（2）
计算节点（N）

图 1

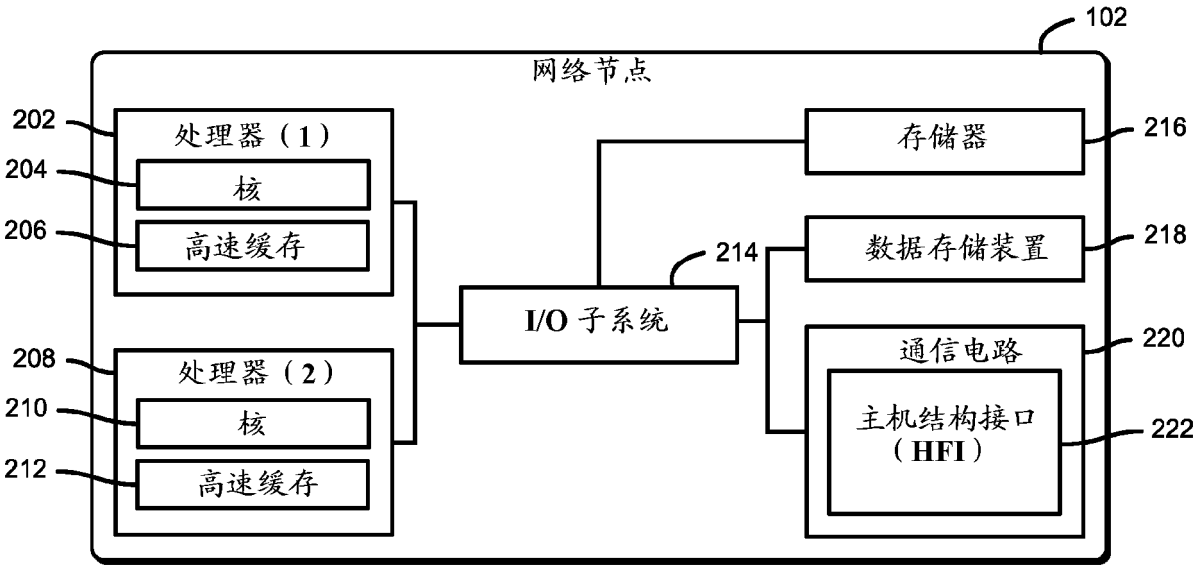

102
网络节点
202
处理器（1）
204
核
206
高速缓存
208
处理器（2）
210
核
212
高速缓存
I/O 子系统
214
存储器
216
数据存储装置
218
通信电路
220
主机结构接口
（HFI）
222

图 2

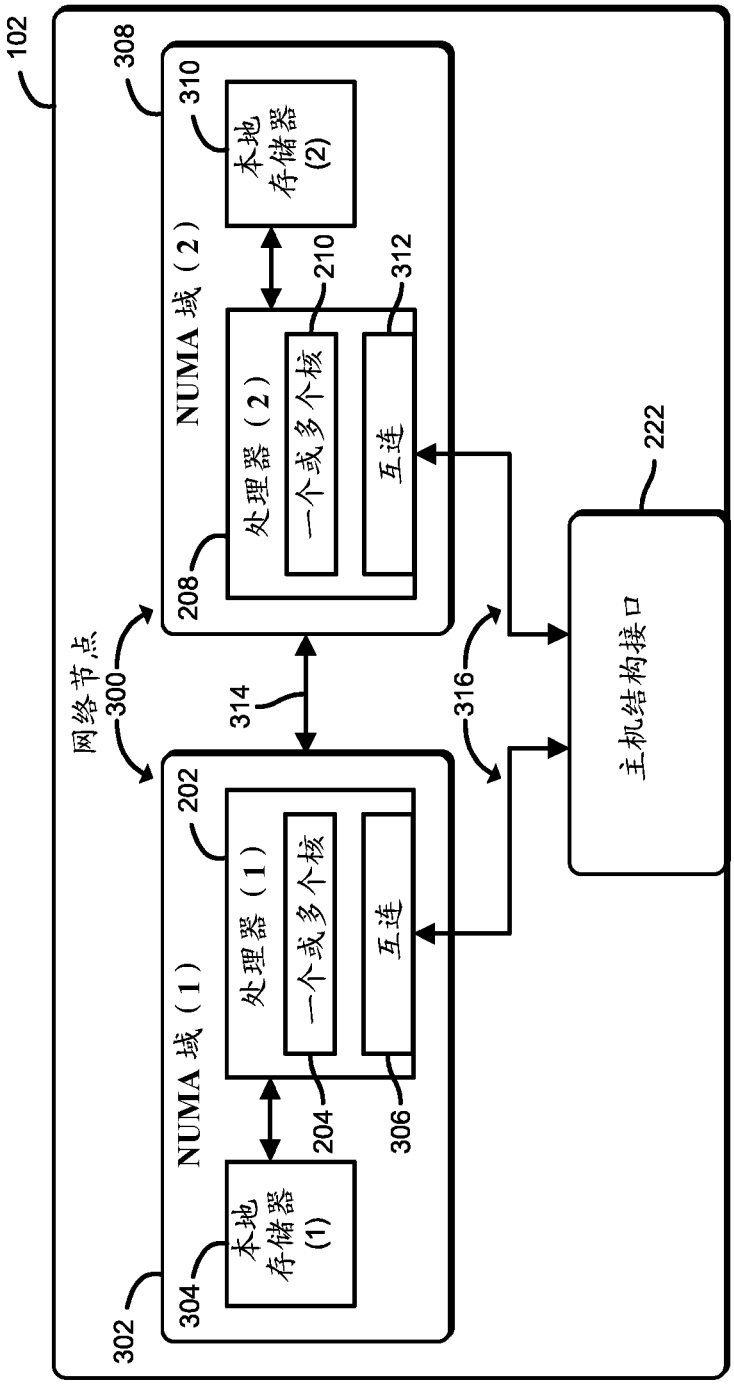

102
308
310
本地存储器(2)
NUMA 域（2）
210
312
处理器（2）
一个或多个核
互连
208
222
主机结构接口
网络节点
300
314
316
202
处理器（1）
一个或多个核
互连
NUMA 域（1）
204
306
本地存储器(1)
304
302

图 3

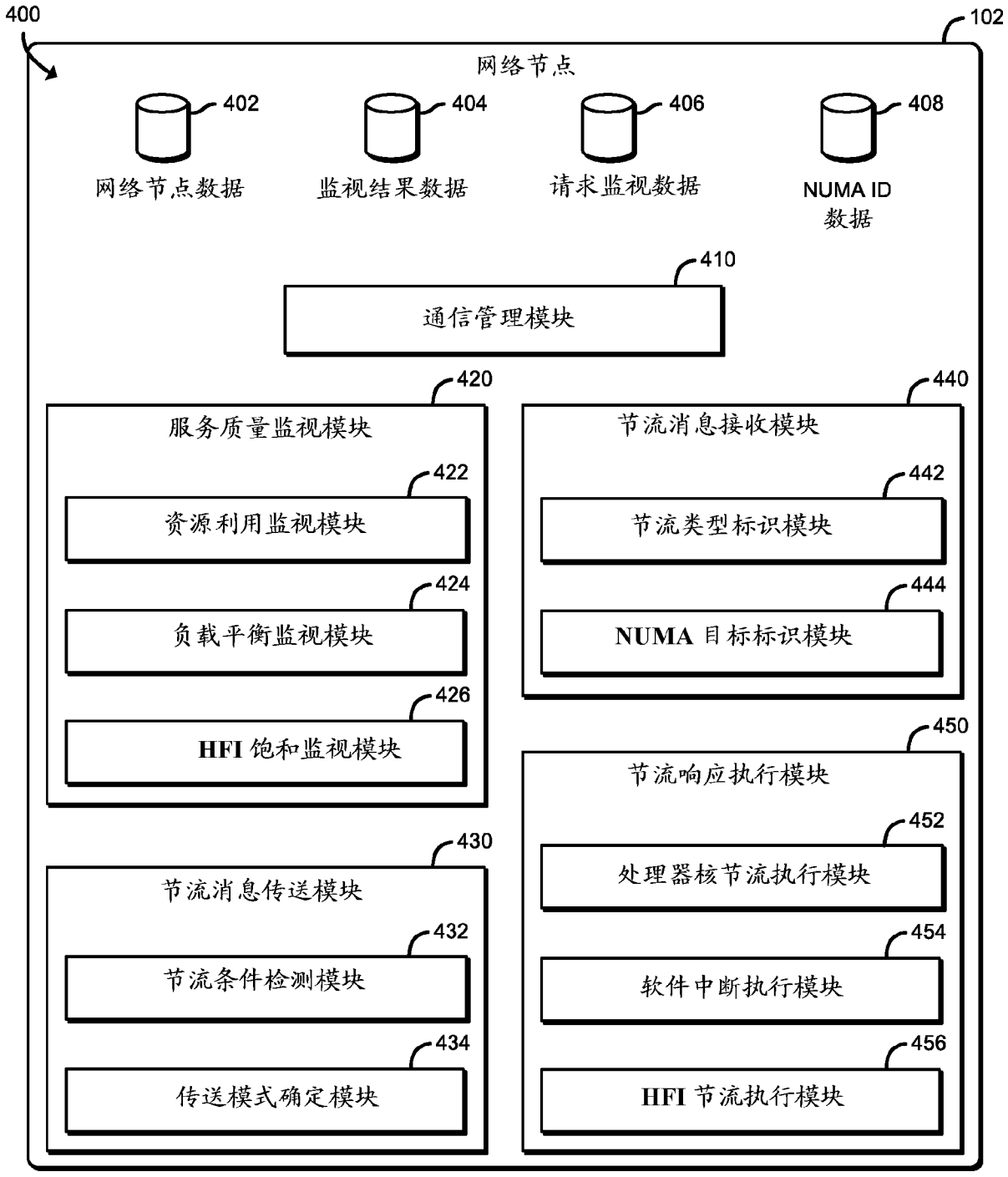

400
102
网络节点
402
网络节点数据
404
监视结果数据
406
请求监视数据
408
NUMA ID
数据
410
通信管理模块
420
服务质量监视模块
422
资源利用监视模块
424
负载平衡监视模块
426
HFI 饱和监视模块
430
节流消息传送模块
432
节流条件检测模块
434
传送模式确定模块
440
节流消息接收模块
442
节流类型标识模块
444
NUMA 目标标识模块
450
节流响应执行模块
452
处理器核节流执行模块
454
软件中断执行模块
456
HFI 节流执行模块

图 4

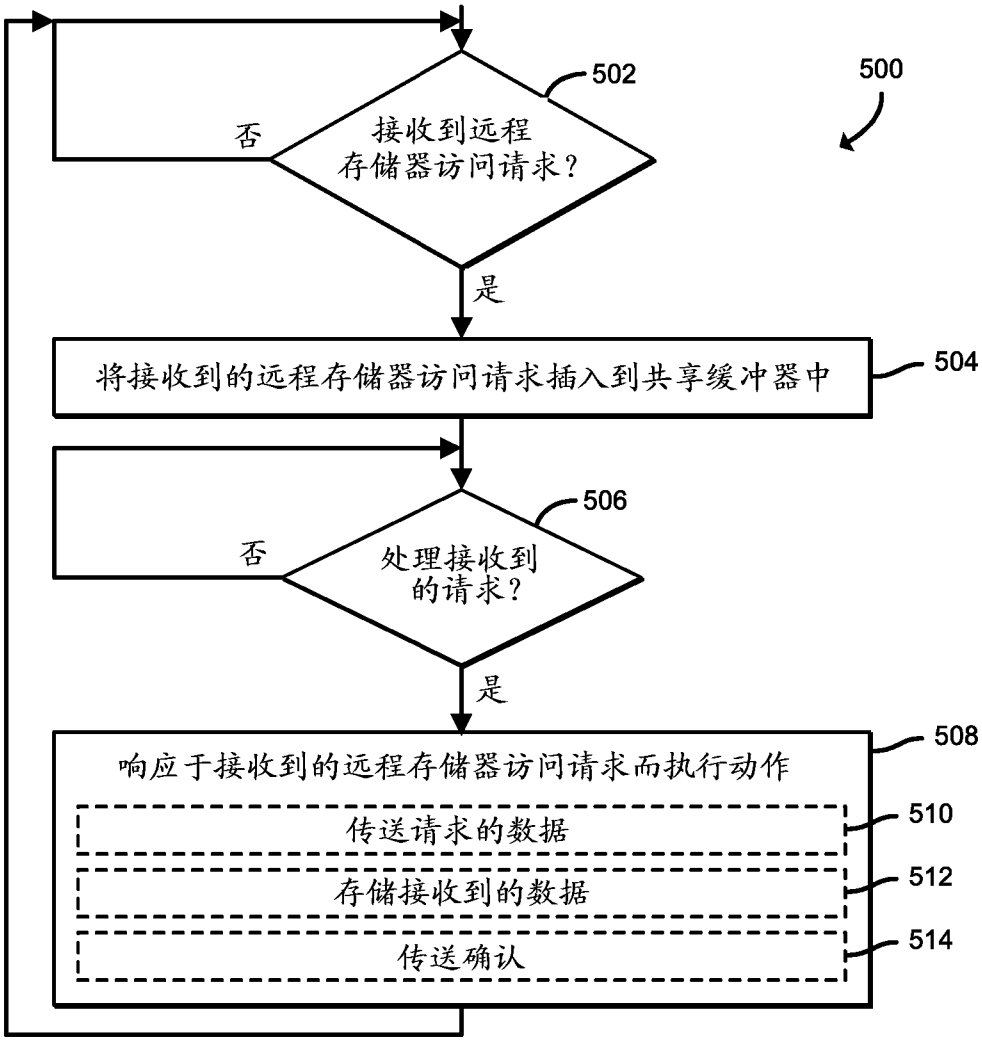

500
502
接收到远程存储器访问请求？
否
是
504
将接收到的远程存储器访问请求插入到共享缓冲器中
506
处理接收到的请求？
否
是
508
响应于接收到的远程存储器访问请求而执行动作
510
传送请求的数据
512
存储接收到的数据
514
传送确认

图 5

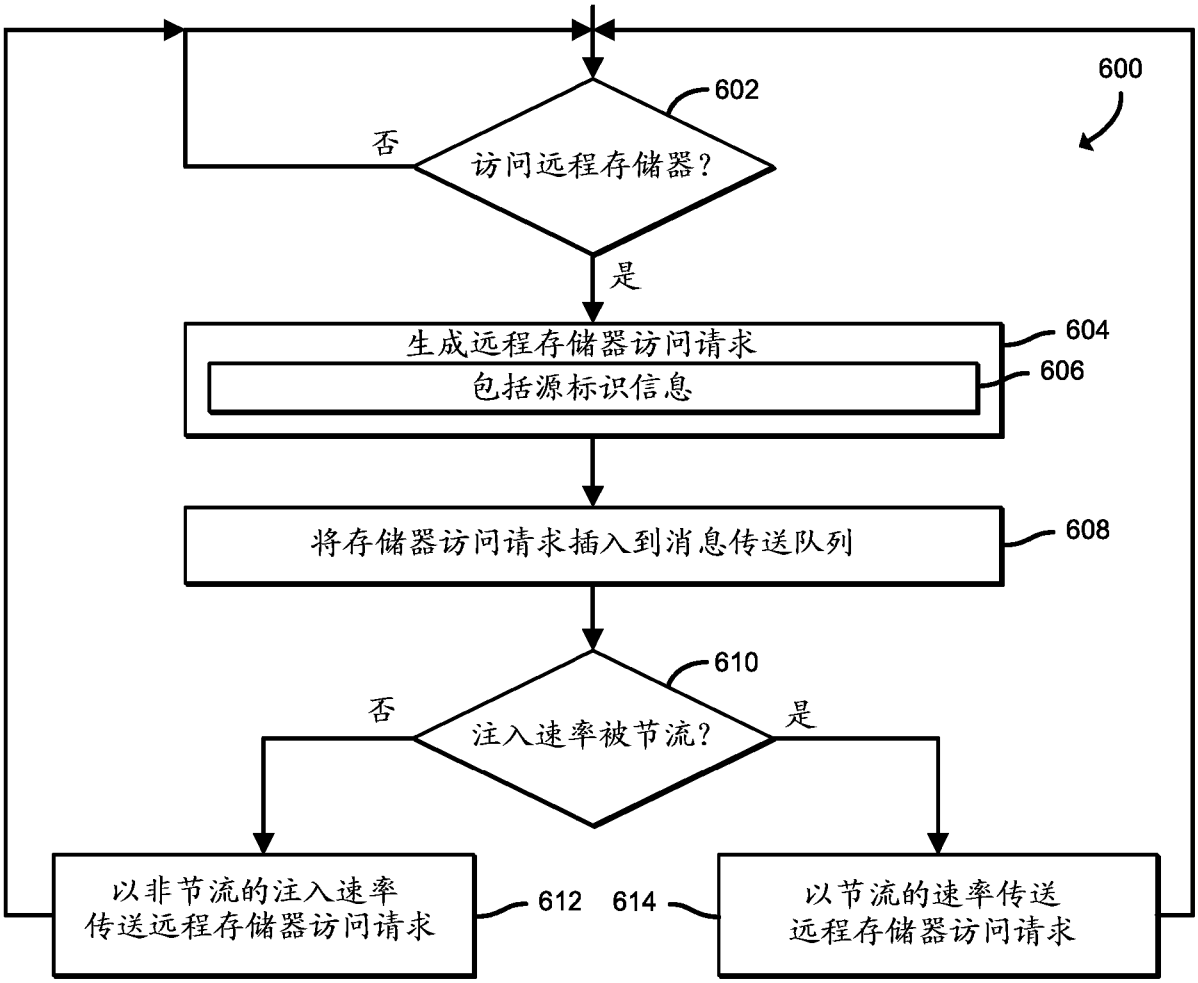

600
602
访问远程存储器?
否
是
生成远程存储器访问请求
包括源标识信息
604
606
将存储器访问请求插入到消息传送队列
608
610
注入速率被节流?
否
是
以非节流的注入速率
传送远程存储器访问请求
612
614
以节流的速率传送
远程存储器访问请求

图 6

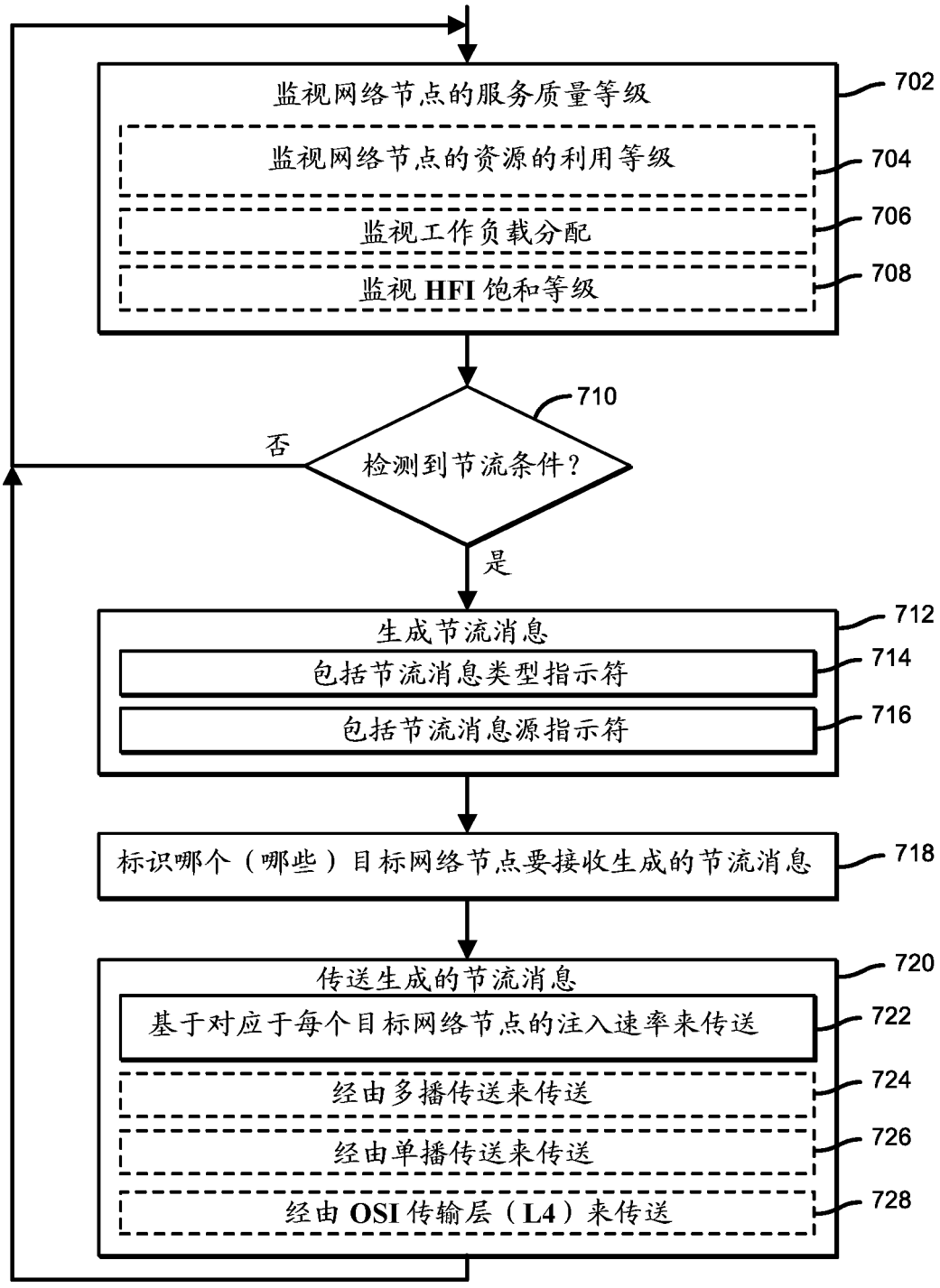

700
监视网络节点的服务质量等级
702
监视网络节点的资源的利用等级
704
监视工作负载分配
706
监视 HFI 饱和等级
708
710
检测到节流条件？
否
是
生成节流消息
712
包括节流消息类型指示符
714
包括节流消息源指示符
716
标识哪个（哪些）目标网络节点要接收生成的节流消息
718
传送生成的节流消息
720
基于对应于每个目标网络节点的注入速率来传送
722
经由多播传送来传送
724
经由单播传送来传送
726
经由 OSI 传输层（L4）来传送
728

图 7

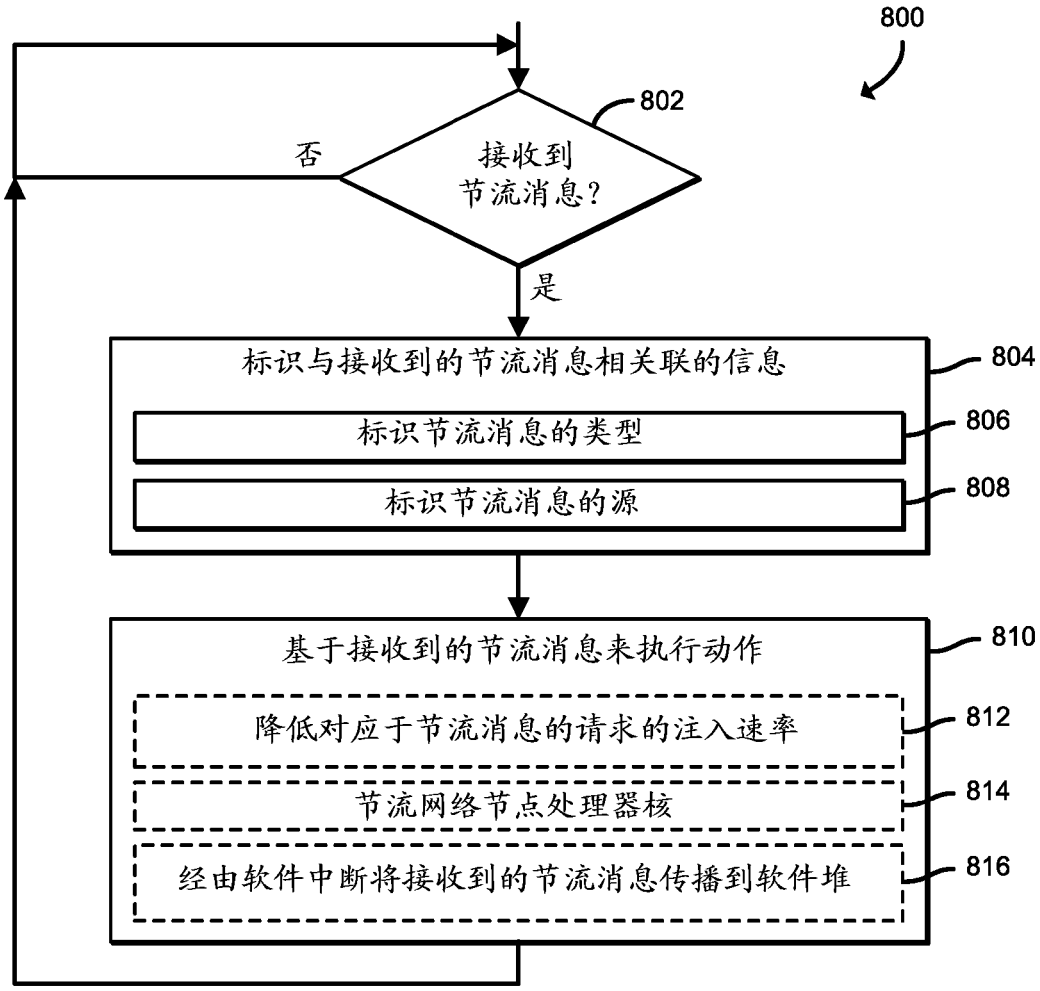

800
802
接收到
节流消息?
否
是
标识与接收到的节流消息相关联的信息
804
标识节流消息的类型
806
标识节流消息的源
808
基于接收到的节流消息来执行动作
810
降低对应于节流消息的请求的注入速率
812
节流网络节点处理器核
814
经由软件中断将接收到的节流消息传播到软件堆
816

图 8