



(12) 发明专利

(10) 授权公告号 CN 1879385 B

(45) 授权公告日 2012.03.21

(21) 申请号 200480033125.9

(22) 申请日 2004.10.29

(30) 优先权数据

10/712,640 2003.11.12 US

(85) PCT申请进入国家阶段日

2006.05.10

(86) PCT申请的申请数据

PCT/US2004/036095 2004.10.29

(87) PCT申请的公布数据

W02005/050949 EN 2005.06.02

(73) 专利权人 英特尔公司

地址 美国加利福尼亚州

(72) 发明人 拉梅什库马·伊利卡尔

(74) 专利代理机构 上海专利商标事务所有限公

司 31100

代理人 钱慰民

(51) Int. Cl.

H04L 29/06 (2006.01)

(56) 对比文件

US 2002/0165992 A1, 2002.11.07, 说明书第
[0011]-[0022], [0036]-[0042] 段、图 4-7.

US 6434620 B1, 2002.08.13, 说明书第 3 栏
第 28-43 行, 第 53 栏第 50 行 - 第 54 栏第 54 行、
图 17.

CN 1326573 A, 2001.12.12, 全文.

审查员 王玉婧

权利要求书 2 页 说明书 3 页 附图 4 页

(54) 发明名称

预取用于网络分组处理的协议控制块的方法
和系统

(57) 摘要

描述了用来为网络分组处理预取协议控制块
的方法和系统。所述方法包括接收分组,把与所述
分组相关联的协议控制块 (PCB) 预取到缓存中,
排队分组以进行处理,以及当处理单元准备好处
理所述分组时,从所述缓存取回所述 PCB。

1. 一种预取用于网络分组处理的协议控制块的方法,包括:
在网络设备上接收分组;
将与所接收的分组相关联的协议控制块 PCB 预取到第一处理单元的缓存中;
将与所接收的分组相关联的头预取到所述第一处理单元的缓存中;
排队所接收的分组以进行处理;以及
当所述第一处理单元准备好处理所接收的分组时,从所述第一处理单元的缓存中取回所接收的分组的所述 PCB 和与所接收的分组相关联的所述头。
2. 如权利要求 1 所述的方法,还包括发送中断以向所述第一处理单元通知接收到所述分组。
3. 如权利要求 1 所述的方法,其中,响应于接收到其连接的上下文被存储在所述 PCB 中的相关联的分组,把所述 PCB 预取到所述第一处理单元的缓存中。
4. 如权利要求 3 所述的方法,还包括把所述分组存储到耦合到所述第一处理单元的存储器中。
5. 如权利要求 1 所述的方法,还包括处理所述分组。
6. 一种预取用于网络分组处理的协议控制块的装置,包括:
接收单元,所述接收单元用来接收分组;
耦合到所述接收单元的预取单元,所述预取单元在所述接收单元接收到所述分组后,把与所述分组相关联的协议控制块 PCB 以及与所述分组相关联的头预取到一处理单元的缓存中,并且排队所述分组以进行处理;以及
耦合到所述预取单元的所述处理单元,所述处理单元用来从所述缓存取回所述 PCB 和所述头,并处理所述分组。
7. 如权利要求 6 所述的装置,其中所述接收单元是网络接口卡。
8. 如权利要求 6 所述的装置,还包括耦合到所述接收单元和所述处理单元的中断单元,所述中断单元从所述接收单元接收中断,并向所述处理单元通知所述分组。
9. 一种预取用于网络分组处理的协议控制块的设备,包括:
用于在网络设备上接收分组的装置;
用于将与所接收的分组相关联的协议控制块 PCB 预取到第一处理单元的缓存中的装置;
用于将与所接收的分组相关联的头预取到所述第一处理单元的缓存中的装置;
用于排队所接收的分组以进行处理的装置;以及
用于当所述第一处理单元准备好处理所接收的分组时,从所述第一处理单元的缓存中取回所接收的分组的所述 PCB 以及与所接收的分组相关联的所述头的装置。
10. 如权利要求 9 所述的设备,其中还包括:用于处理所述分组的装置。
11. 如权利要求 9 所述的设备,其中还包括:用于发送中断以向所述第一处理单元通知接收到所述分组的装置。
12. 如权利要求 9 所述的设备,其中还包括:用于把所述分组存储在耦合到所述第一处理单元的存储器中的装置。
13. 一种预取用于网络分组处理的协议控制块的系统,包括:
接收单元,所述接收单元用来接收分组;

耦合到所述接收单元的存储器,所述存储器用来存储接收到的分组;

耦合到所述存储器的存储器控制器,所述存储器控制器用来管理所述存储器;

耦合到所述接收单元的预取单元,所述预取单元在所述接收单元接收到所述分组后,把与所述分组相关联的协议控制块 PCB 以及与所述分组相关联的头预取到缓存中,并且排队所述分组以进行处理;以及

处理单元,所述处理单元用来从所述缓存取回所述 PCB 和所述头,并处理所述分组。

14. 如权利要求 13 所述的系统,其中所述接收单元是网络接口卡。

15. 如权利要求 13 所述的系统,还包括耦合到所述接收单元和所述处理单元的中断单元,所述中断单元从所述接收单元接收中断,并向所述处理单元通知所述分组。

预取用于网络分组处理的协议控制块的方法和系统

[0001] 背景

1. 技术领域

[0002] 本发明的实施方案涉及网络分组处理领域,更具体地说,涉及预取用于网络分组处理的协议控制块。

[0003] 2. 背景信息和相关技术的描述

[0004] 当分组(packet)到达网络设备时,网络接口卡(NIC)接受(take)所述分组并将其存储在主存储器中。接着NIC可以发送中断以向中央处理器(CPU)通知该分组。然后,中断单元可以检查中断的目的地(destination),禁止(disable)来自NIC的后续中断,发起(initiate)软件中断,并且排队所述分组以进行处理。当处理单元准备好处理分组的时候,识别分组所属的连接。这可以包括取(fetch)与分组相关联的协议控制块(PCB)。取了PCB后,CPU可以开始处理分组。当处理单元准备好处理分组时,取PCB所发生的存储器延迟(latency)会降低网络设备的性能。随着网络速度的增加,对于性能和吞吐量,此存储器延迟成为越来越大的问题。

附图说明

[0005] 可以通过参考以下描述和附图来最好地理解本发明,所述附图用来图示说明本发明的实施方案。在附图中:

[0006] 图1是图示结合本发明的系统的一个一般性实施方案的框图。

[0007] 图2是根据本发明的一个实施方案,示出结合本发明的示例性系统的框图。

[0008] 图3是根据本发明的一个实施方案,示出方法的流程图。

[0009] 图4是可以实践被图示说明的发明的某些方面的合适计算环境的框图。

具体实施方式

[0010] 描述了预取协议控制块用于网络分组处理的系统和方法的实施方案。在下面的描述中,阐述了很多具体的细节。然而,可以理解,无需这些具体的细节可以实现本发明。此外,公知的电路、结构和技术没有详细示出,以免模糊对本发明的理解。

[0011] 在整篇说明书中提及“一个实施方案”或“实施方案”意味着关于该实施方案描述的具体特征、结构或特性被包括在本发明的至少一个实施方案中。因此,“在一个实施方案中”或“在实施方案中”在整篇说明书中不同地方的出现不一定全是指同一实施方案。此外,所述具体特征、结构或特性可以以任何适当的方式结合在一个或更多个实施方案中。

[0012] 参考图1,根据本发明的一个实施方案,框图示出了系统100。本领域的普通技术人员会理解,系统100可以包括比图1示出的那些组件更多的组件。但是,对于实践本发明,不需要为了公开说明性的实施方案而示出所有这些一般性的常规组件。

[0013] 系统100包括接收单元102,以接收来自网络的分组。在一个实施方案中,分组可以根据传输控制协议(TCP/IP)规范进行传送。特定的一连串分组可以被称为“连接”或“分

组流 (flow)”。连接的上下文 (context) 可以存储在被称为协议控制块 (PCB) 的结构中。此上下文可以由连接的源 IP 地址、目的 IP 地址、源端口、目的端口和 / 或协议类型唯一地确定。对于每个在接收单元 102 接收到的分组, 与所述分组相关联的 PCB 可能需要从存储器中取回 (retrieved)。

[0014] 针对每个被发送和接收的分组而从存储器访问 PCB 的操作具有相关联的存储器延迟和带宽问题。为了减少这些问题, 预取单元 104 把与被接收的分组相关联的 PCB 取入处理单元 106 的缓存 (cache) 108 中。然后, 排队所述分组以进行处理。当处理单元 106 准备好处理分组时, PCB 可以从其缓存 108 中被取回。因此, 当处理单元准备好处理分组时, 取 PCB 的存储器延迟被减少。

[0015] 在一个实施方案中, 预取单元 104 还将分组头 (header) 信息预取到缓存 108 中。当处理单元 106 准备好处理分组时, 分组的头信息可以从缓存 108 取回, 然后分组可以被处理。

[0016] 图 2 根据本发明的一个实施方案, 示出了结合本发明的示例性系统。在此实施方案中, 网络接口卡 (NIC) 202 接收来自网络的分组。分组通过存储器控制器 212 存储在主存储器 214 中。NIC 202 发送中断, 以向处理单元 (例如 206、208 或 210) 通知该分组。例如中断服务循环 (rotate) (ISR) 单元的中断单元 204 检查中断的目的地址, 禁止来自 NIC 的后续中断, 发起软件中断, 并且排队分组以进行处理。在一个实施方案中, 通过排队延期过程调用 (deferred procedure call, DPC) 来排队分组, 以进行处理。这时, 可以发起与分组相关联的 PCB 预取。还可以发起分组头的预取。PCB 和分组头可以被预取到缓存 (例如 216、218、或 220) 中。预取可以在硬件或软件中完成。当处理单元 (例如 206、208、或 210) 准备好处理分组时, 处理单元可以从其缓存中取出 PCB 和分组头。然后, 分组可以被处理。接着, 处理单元可以使能 (enable) 来自 NIC 的中断。

[0017] 在一个实施方案中, 预取也可以用在发送侧, 以减少存储器延迟。当分组被排队以传输到网络外时, PCB 可以被预取。用户空间中的套接字 (socket) 接口与内核 (kernel) 空间中的协议栈处理间经常有延迟。所以, 为分组发起发送请求时, 与分组相关联的 PCB 可以被预取。当内核准备好处理用于传输的分组时, PCB 已经被预取并且准备好进行处理。这减少了发送侧的存储器延迟。

[0018] 图 3 根据本发明的一个实施方案示出了一种方法。在 300, 接收到分组。在 302, 与所述分组相关联的 PCB 被预取到缓存中。在一个实施方案中, 分组头也被预取到缓存中。在 304, 排队所述分组用于处理。在 306, 当处理单元准备好处理所述分组时, 从缓存中取回 PCB。在一个实施方案中, 分组的头也从缓存中取回。然后, 分组可以被处理。

[0019] 图 4 是可以实践所说明的发明的某些方面的合适计算环境的框图。在一个实施方案中, 上述方法可以在具有组件 402-412 的计算机系统 400 上实现, 所述组件包括处理器 402、存储器 404、输入 / 输出器件 406、数据存储设备 412 和网络接口 410, 这些组件通过总线 408 互相耦合。这些组件执行它们的技术领域中公知的常规功能, 并且提供实现系统 100 的装置。这些组件一起代表硬件系统的一个宽泛的类别, 包括但不限于通用计算机系统和专用的分组转发设备。应该理解, 计算机系统 400 的各种组件可以重新安排, 并且本发明的某些实现可能并不需要也并不包括上面所有的组件。此外, 额外的组件可以被包括在系统 400 中, 例如额外的处理器 (例如数字信号处理器)、存储设备、存储器以及网络或通信接

口。

[0020] 正如本领域技术人员将会理解的那样,用于实现本发明的方法的实施方案的内容,例如计算机程序指令,可以由任何能够存储系统 100 可访问的数据的机器可读介质来提供,所述机器可读介质可以作为存储器的一部分或是存储器的附加,包括但不限于磁带(cartridge)、磁性盒带(magnetic cassette)、闪存存储卡、数字视频卡、随机访问存储器(RAM)、只读存储器(ROM)等等。在此方面,系统 100 被装配为以本领域公知的方式与这些机器可读介质进行通信。

[0021] 本领域技术人员还将理解,用于实现本发明的方法的实施方案的内容可以从任何能够存储所述内容并向系统 100 传送这些内容的外部设备提供给系统 100。例如,在本发明的一个实施方案中,系统 100 可以连接到网络,并且所述内容可以存储在网络中的任何设备上。

[0022] 尽管关于几个实施方案描述了本发明,但是本领域的技术人员将认识到本发明并不限于所描述的实施方案,而是可以在所附权利要求书的精神和范围内以修改和改变来实现本发明。因此说明书应该被认为是示例性而非限制性的。

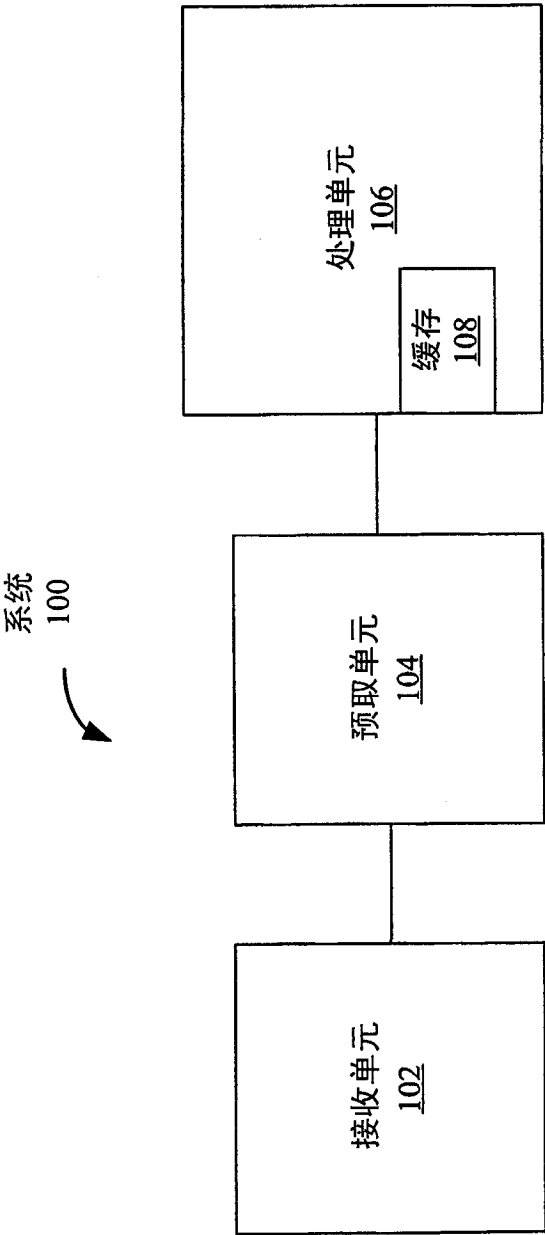


图 1

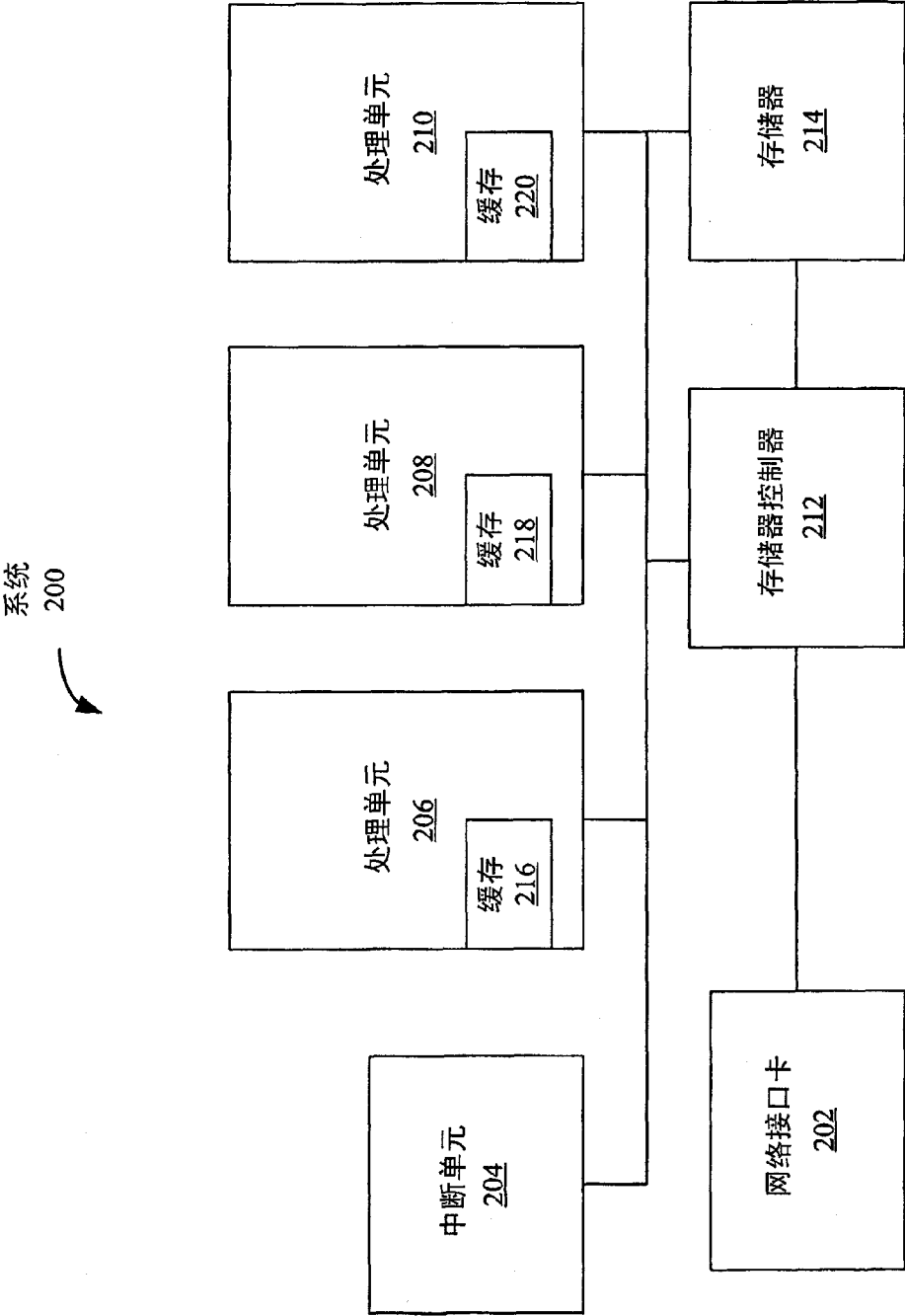


图 2

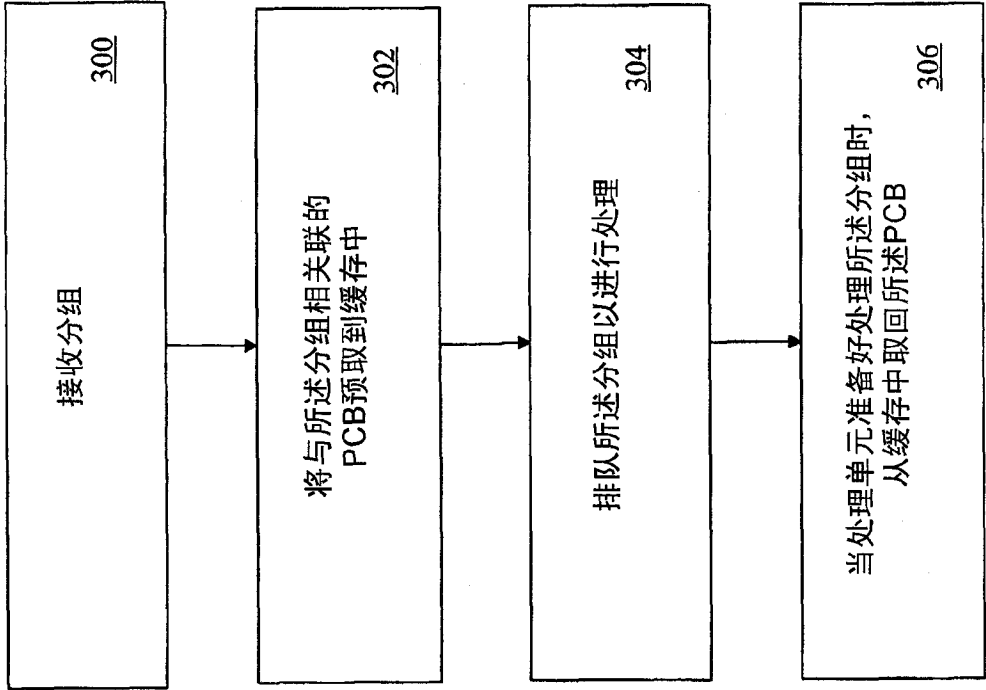


图 3

计算机系统
400

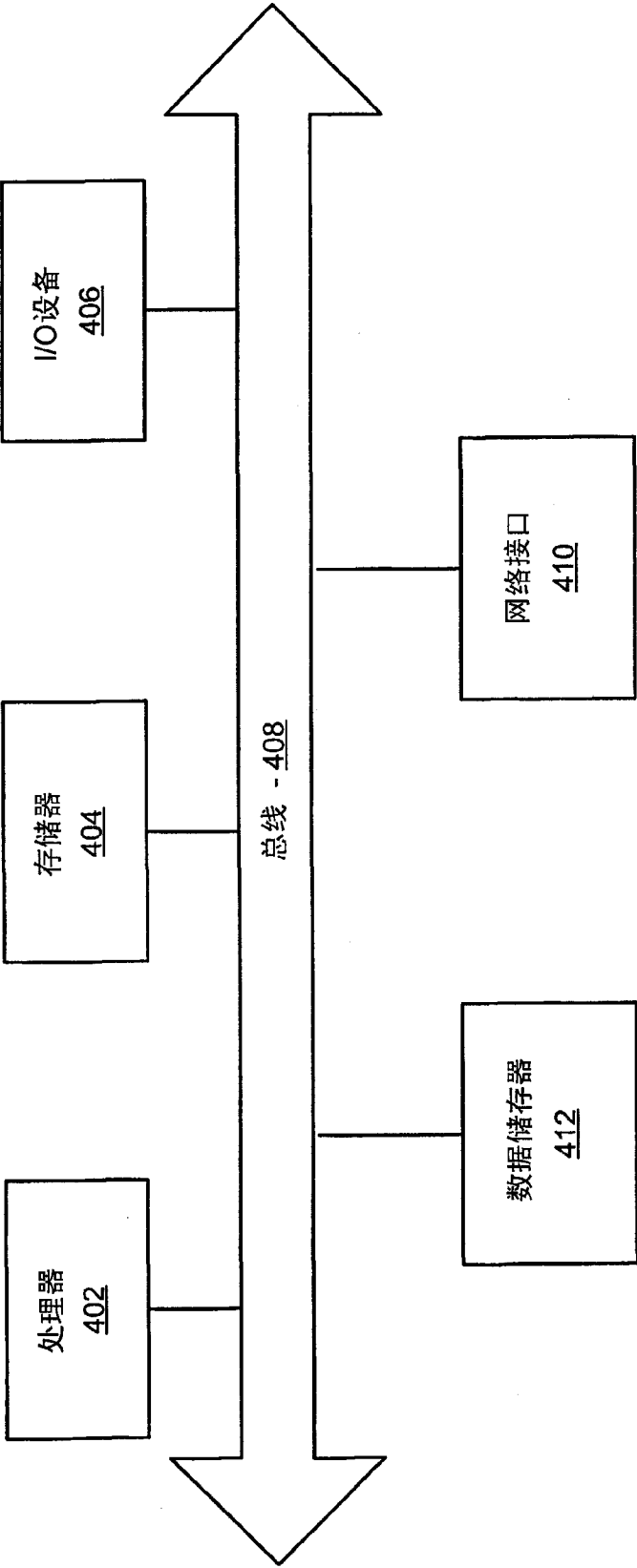


图 4