



(19)中華民國智慧財產局

(12)發明說明書公開本 (11)公開編號：TW 202415044 A

(43)公開日：中華民國 113 (2024) 年 04 月 01 日

(21)申請案號：112131207 (22)申請日：中華民國 112 (2023) 年 08 月 18 日
(51)Int. Cl. : H04L67/01 (2022.01) H04L69/085 (2022.01)
(30)優先權：2022/08/19 美國 63/373,016
2023/05/19 美國 63/503,349
(71)申請人：美商特斯拉公司 (美國) TESLA, INC. (US)
美國
(72)發明人：奎內爾 艾瑞克 QUINNELL, ERIC C. (US)；威廉斯 道格拉斯 WILLIAMS,
DOUGLAS R. (US)；熊 克里斯多夫 HSIONG, CHRISTOPHER (US)；納瓦烏塔
多 赫拉多 NAVARRO HURTADO, GERARDO (US)
(74)代理人：林志剛
申請實體審查：無 申請專利範圍項數：60 項 圖式數：13 共 97 頁

(54)名稱
用於乙太網的傳輸協議
(57)摘要

本公開涉及用於在基於乙太網的網絡中使用傳輸層進行通信而無需軟件控制機制的輔助的系統和方法。在一些實施例中，第一節點被配置為根據第一節點的狀態機硬件來打開和關閉與基於乙太網的網絡中的第二節點的鏈路。第一節點可以包括硬件鏈路定時器，其被配置為確定是否重放鏈路。第一節點可以包括硬件重放架構，其被配置為以僅硬件重放鏈路。

The present disclosure relates to systems and methods for communicating in an Ethernet-based network using a transport layer without assistance of software-controlled mechanisms. In some embodiments, a first node is configured to open and close a link with a second in the Ethernet-based network according to a state machine hardware of the first node. The first node can include a hardware link timer configured to determine whether to replay the link. The first node can include a hardware replay architecture configured to replay the link in hardware only.

指定代表圖：

OSI層	示例協議 (TCP/IP)	TCP/IP 實施方式
第7層 應用	HTTP, Telnet, FTP	軟件
第6層 呈現	JPEG, PNG, MPEG	
第5層 會話	NFS, SQL	
第4層 傳輸	TCP, UDP	
第3層 網絡	IPv4/IPv6	
第2層 數據鏈路	乙太網幀、MAC地址、VLAN	硬件
第1層 物理	數據編碼、物理規範	

【圖 1A】

【發明摘要】

【中文發明名稱】

用於乙太網的傳輸協議

【英文發明名稱】

TRANSPORT PROTOCOL FOR ETHERNET

【中文】

本公開涉及用於在基於乙太網的網絡中使用傳輸層進行通信而無需軟件控制機制的輔助的系統和方法。在一些實施例中，第一節點被配置為根據第一節點的狀態機硬件來打開和關閉與基於乙太網的網絡中的第二節點的鏈路。第一節點可以包括硬件鏈路定時器，其被配置為確定是否重放鏈路。第一節點可以包括硬件重放架構，其被配置為以僅硬件重放鏈路。

【英文】

The present disclosure relates to systems and methods for communicating in an Ethernet-based network using a transport layer without assistance of software-controlled mechanisms. In some embodiments, a first node is configured to open and close a link with a second in the Ethernet-based network according to a state machine hardware of the first node. The first node can include a hardware link timer configured to determine whether to replay the link. The first node can include a hardware replay architecture configured to replay the link in hardware only.

【指定代表圖】圖 1A
【代表圖之符號簡單說明】無

【特徵化學式】無

【發明說明書】

【中文發明名稱】

用於乙太網的傳輸協議

【英文發明名稱】

TRANSPORT PROTOCOL FOR ETHERNET

【技術領域】

本公開涉及促進在網絡上通信的系統和方法。更特別地，本公開的實施例涉及可使用硬件實現的流控制協議，用於在基於乙太網的網絡上通信。

優先權申請的交叉引用

本申請為2022年8月19日提交的題為“TRANSPORT PROTOCOL FOR ETHERNET”的美國臨時專利申請第63/373,016號的非臨時申請，並要求其優先權，該申請的技術公開內容通過引用以其整體並且出於所有目的併入於此。本申請是2023年5月19日提交的題為“TRANSPORT PROTOCOL FOR ETHERNET”的美國臨時專利申請第63/503,349號的非臨時申請，並要求其優先權，該申請的技術公開內容通過引用以其整體並且出於所有目的併入於此。

【先前技術】

電氣和電子工程師協會(IEEE)已經為局域網(LAN)提供了各種標準，統稱為IEEE 802，其包括通常稱為乙太網的IEEE 802.3標準。IEEE 802.3乙太網標準具有用於物理介質接口(乙太網線纜、光纖、背板等)的規範，但是不用於通信的流控制。諸如TCP/IP、RoCE或InfiniBand的協議可以加速結構流控制。TCP/IP協議一般具有通常處於毫秒量級的時延，而RoCE或InfiniBand具有無損和擴展規範，這可能過度約束了系統。

隨著高性能計算(HPC)和人工智能(AI)訓練數據中心變得越來越普遍，期望具有高帶寬、低時延、有損縮放彈性、分布式控制、和盡可能少的軟件開銷的通信網絡結構。照此，可能期望開發可在基於有損乙太網的網絡上操作的網絡流控制協議，而很少有或沒有中央處理單元(CPU)的參與，同時實現比現有的基於乙太網的網絡更低的時延。

【發明內容】

本公開的系統、方法和設備每個具有若干創新實施例，其中沒有一個單獨負責本文中公開的所有期望屬性。本說明書中描述的主題的一個或多個實施方式的細節在隨附圖式和下面的描述中闡述。

在一些方面中，本文中描述的技術涉及一種用於基於乙太網的通信的第一節點，所述第一節點包括：一個或多個處理器，其被配置為實現傳輸層僅硬件乙太網協議。

在一些方面中，本文中描述的技術涉及第一節點，其中乙太網協議是有損的。

在一些方面中，本文中描述的技術涉及第一節點，其中所述一個或多個處理器進一步被配置為實現硬件重放架構，以重放在第一鏈路上傳輸至第二節點的分組，其中所述分組存儲在第一節點的本地存儲裝置中，並且其中用於重放的分組的次序在鏈表中指定。

在一些方面中，本文中描述的技術涉及第一節點，其中所述第一節點被配置為以一位數微秒時延向第二節點傳輸分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述一個或多個處理器被配置為實現狀態機，該狀態機被配置為：在第一節點和第二節點之間的鏈路打開的打開狀態下操作；從打開狀態轉換到中間關閉狀態；以及響應於從第二節點接收到關閉確認，從中間關閉狀態轉換到關閉狀態以關閉鏈路。

在一些方面中，本文中描述的技術涉及第一節點，進一步包括乙太網端口。

在一些方面中，本文中描述的技術涉及第一節點，其中所述一個或多個處理器被配置為基於存儲在先進先出(FIFO)存儲器中的與所述鏈路相關聯的時序和狀態信息，確定在第一節點和第二節點之間的鏈路上重放分組，其中根據與多個鏈路相關聯的硬件鏈路定時器的滴答來訪問FIFO存儲器的條目。

在一些方面中，本文中描述的技術涉及一種用於基於乙太網的通信的第一節點，所述第一節點包括：一個或多個處理器，其被配置為實現第2層僅硬件乙太網協議。

在一些方面中，本文中描述的技術涉及第一節點，其中所述一個或多個處理器包括僅硬件架構，所述僅硬件架構被配置為重放在第一鏈路上傳輸到第二節點的分組。

在一些方面中，本文中描述的技術涉及第一節點，一個或多個處理器進一步被配置為基於存儲在先進先出(FIFO)存儲器中的與關聯於第一節點的鏈路相關聯的時序和狀態信息來確定在所述鏈路上重放分組，基於與多個鏈路相關聯的定時器的滴答來訪問所述FIFO存儲器。

在一些方面中，本文中描述的技術涉及第一節點，其中所述第一節點被配置為打開和關閉與基於乙太網的網絡中的第二節點的鏈路，第一節點包括：狀態機硬件，其被配置為：在第一節點和第二節點之間的鏈路打開的打開狀態下操作；從打開狀態轉換到中間關閉狀態；以及響應於從第二節點接收到關閉確認，從中間關閉狀態轉換到關閉狀態以關閉鏈路，其中所述第一節點被配置為在有損網絡中操作。

在一些方面中，本文中描述的技術涉及第一節點，其中所述狀態機硬件以僅硬件實現傳輸層的流控制協議。

在一些方面中，本文中描述的技術涉及第一節點，其中與流控制協議相關聯的時延小於10微秒。

在一些方面中，本文中描述的技術涉及第一節點，其

中所述狀態機硬件被配置為：從關閉狀態轉換到中間打開狀態；並且從中間打開狀態轉換到打開狀態。

在一些方面中，本文中描述的技術涉及第一節點，其中所述狀態機硬件響應於向第二節點傳輸關閉鏈路的請求或從第二節點接收關閉鏈路的請求，從打開狀態轉換到中間關閉狀態。

在一些方面中，本文中描述的技術涉及第一節點，其中所述狀態機硬件響應於向第二節點傳輸關閉鏈路的確認，從中間關閉狀態轉換到關閉狀態。

在一些方面中，本文中描述的技術涉及第一節點，其中所述狀態機硬件從中間關閉狀態轉換到關閉狀態，無需等待一段時間。

在一些方面中，本文中描述的技術涉及第一節點，其中在打開狀態下，第一節點不重傳分組，直到從第二節點接收到分組的未確認，或者預確定的超時時段到期而未接收到分組的未確認為止。

在一些方面中，本文中描述的技術涉及第一節點，其中在打開狀態下，第一節點不暫停地傳輸最多 N 個分組，並且其中 N 受分配給第一節點的物理存儲器的大小限制。

在一些方面中，本文中描述的技術涉及第一節點，進一步包括：與多個鏈路相關聯的硬件鏈路定時器；以及被配置為以僅硬件重放分組的硬件重放架構。

在一些方面中，本文中描述的技術涉及第一節點，包括：硬件重放架構，其被配置為重放在第一鏈路上使用乙

太網協議傳輸至第二節點的分組，其中所述硬件重放架構包括：本地存儲裝置，其被配置為存儲包括分組的鏈表，其中所述鏈表維護分組傳輸至第二節點的次序；以及邏輯電路，其被配置為：響應於(a)從第二節點接收到第一分組的未確認或者(b)與第一分組相關聯的超時中的至少一個，確定重放分組中的第一分組；以及響應於從第二節點接收到第二分組的確認，退出分組中的第二分組，其中乙太網協議是有損的。

在一些方面中，本文中描述的技術涉及第一節點，其中所述邏輯電路包括多個流水線級，並且其中所述邏輯電路確定在所述多個流水線級中的第一流水線級處處理與第一節點和第二節點之間的第一鏈路而非第二鏈路相關聯的數據。

在一些方面中，本文中描述的技術涉及第一節點，其中所述邏輯電路確定在所述多個流水線級中的第二流水線級重放第一分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述邏輯電路在所述多個流水線級中的第二流水線級處，基於由鏈表維護的分組的次序，確定重放分組中的第三分組和分組中的第一分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述邏輯電路基於鏈路指針確定處理與第一鏈路而非第二鏈路相關聯的數據，並且其中所述邏輯電路更新鏈路指針以指向所述多個流水線級中的第三流水線級處的第二鏈

路。

在一些方面中，本文中描述的技術涉及第一節點，其中第一節點和第二節點在基於乙太網的網絡中，並且其中第一節點通過乙太網交換機與第二節點通信。

在一些方面中，本文中描述的技術涉及第一節點，其中所述第一節點包括網絡接口處理器(NIP)和高帶寬存儲器(HBM)，其中HBM的帶寬至少為一千兆字節。

在一些方面中，本文中描述的技術涉及一種用於基於乙太網的通信的第一節點，所述第一節點包括：一個或多個處理器，其被配置為實現傳輸層僅硬件乙太網協議，其中所述傳輸層僅硬件乙太網協議是有損的，並且其中所述一個或多個處理器包括硬件重放架構，其被配置為重放在傳輸層僅硬件乙太網協議下傳輸的分組。

在一些方面中，本文中描述的技術涉及第一節點，其中硬件重放架構包括：本地存儲裝置，其被配置為存儲在傳輸層僅硬件乙太網協議下傳輸的分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述硬件重放架構包括：存儲在本地存儲裝置中的鏈表，並且所述鏈表被配置為跟蹤用於傳輸至另一節點的分組的次序，其中所述鏈表的每個元素對應於存儲在本地存儲裝置中的每個分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述硬件重放架構被配置為以對應於鏈表的次序傳輸分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述硬件重放架構被配置為存儲：第一指針，其被配置為指向鏈表的第一元素，其中所述第一指針指示不重放對應於鏈表的第一元素的分組中的第一分組；以及第二指針，其被配置為指向鏈表的第二元素，其中所述第二指針指示重放對應於鏈表的第二元素的分組中的第二分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述硬件重放架構根據用於傳輸的分組的次序重放第二分組和第二分組之後的一個或多個分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述硬件重放架構使本地存儲裝置根據用於傳輸的分組的次序丟棄第一分組和第二分組之前的一個或多個分組。

在一些方面中，本文中描述的技術涉及一種在第一節點處實現的計算機實現的方法，用於使用乙太網協議重放在第一鏈路上傳輸至第二節點的分組，該計算機實現的方法包括：存儲包括分組的鏈表，其中該鏈表維護分組傳輸至第二節點的次序；響應於(a)從第二節點接收到第一分組的未確認或者(b)與第一分組相關聯的超時中的至少一個，確定重放分組中的第一分組；以及響應於從第二節點接收到第二分組的確認，退出分組中的第二分組，其中乙太網協議是有損的。

在一些方面中，本文中描述的技術涉及計算機實現的方法，其中第一節點包括硬件重放架構，所述硬件重放架構包括多個流水線級，並且其中硬件重放架構確定在多個

流水線級中的第一流水線級處處理與第一鏈路而非第二鏈路相關聯的數據。

在一些方面中，本文中描述的技術涉及計算機實現的方法，其中硬件重放架構確定在多個流水線級中的第二流水線級處重放第一分組。

在一些方面中，本文中描述的技術涉及計算機實現的方法，其中所述硬件重放架構基於在所述多個流水線級中的第二流水線級的鏈表所維護的分組的次序，確定重放分組中的第三分組和分組中的第一分組。

在一些方面中，本文中描述的技術涉及計算機實現的方法，其中第一節點和第二節點在基於乙太網的網絡中，並且其中第一節點通過乙太網交換機與第二節點通信。

在一些方面中，本文中描述的技術涉及計算機實現的方法，其中第一節點包括網絡接口處理器(NIP)和高帶寬存儲器(HBM)，其中HBM的帶寬至少為一千兆字節。

在一些方面中，本文中描述的技術涉及一種用於在基於乙太網的網絡中傳輸分組的第一節點，所述第一節點包括：一個或多個處理器，所述一個或多個處理器包括：先進先出(FIFO)存儲器，其被配置為存儲與多個鏈路相關聯的時序和狀態信息，其中所述第一節點被配置為使用乙太網協議在多個鏈路上向一個或多個其他節點傳輸分組；定時器，其被配置為根據時間段滴答，其中定時器與多個鏈路相關聯；以及邏輯電路，其被配置為：基於定時器上的相應滴答來訪問FIFO存儲器的條目；以及基於與多個鏈路

中的第一鏈路相關聯的時序和狀態信息，確定重放與第一鏈路相關聯的至少一個分組，其中乙太網協議是有損的。

在一些方面中，本文中描述的技術涉及第一節點，其中所述邏輯電路被配置為以輪詢方式訪問FIFO存儲器的條目。

在一些方面中，本文中描述的技術涉及第一節點，其中所述定時器被配置為基於與FIFO存儲器的條目相關聯的活動鏈路的數量來調整所述時間段，其中所述活動鏈路被包括在所述多個鏈路中。

在一些方面中，本文中描述的技術涉及第一節點，其中所述邏輯電路被配置為基於與所述多個鏈路中的第二鏈路相關聯的時序和狀態信息，確定退出與第二鏈路相關聯的分組。

在一些方面中，本文中描述的技術涉及第一節點，其中與第二鏈路相關聯的分組存儲在第一節點的本地存儲裝置中，並且其中邏輯電路使本地存儲裝置響應於確定退出與第二鏈路相關聯的分組而丟棄與第二鏈路相關聯的分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述邏輯電路被配置為基於與所述多個鏈路中的第二鏈路相關聯的時序和狀態信息來確定關閉第二鏈路。

在一些方面中，本文中描述的技術涉及第一節點，其中與所述多個鏈路中的第一鏈路相關聯的時序和狀態信息指示第一節點在用於重放分組的閾值持續時間內未接收到

對接收與第一鏈路相關聯的至少一個分組的確認。

在一些方面中，本文中描述的技術涉及一種用於基於乙太網的通信的第一節點，所述第一節點包括：一個或多個處理器，其被配置為實現傳輸層僅硬件乙太網協議，其中所述傳輸層僅硬件乙太網協議是有損的，並且其中所述一個或多個處理器包括硬件鏈路定時器，其被配置為確定在傳輸層僅硬件乙太網協議下傳輸的分組以進行重放。

在一些方面中，本文中描述的技術涉及第一節點，其中第一節點根據傳輸層僅硬件乙太網協議，在第一鏈路上傳輸第一多個分組並在第二鏈路上傳輸第二多個分組，並且其中硬件鏈路定時器包括：先進先出(FIFO)存儲器，其被配置為在FIFO存儲器的第一條目中存儲與第一鏈路相關聯的時序和狀態信息，並在FIFO存儲器的第二條目中存儲與第二鏈路相關聯的時序和狀態信息。

在一些方面中，本文中描述的技術涉及第一節點，其中硬件鏈路定時器包括與根據時間段滴答的多個鏈路相關聯的定時器，其中硬件鏈路定時器以定時器的輪詢方式滴答來訪問FIFO存儲器的條目，其中所述條目包括第一條目和第二條目。

在一些方面中，本文中描述的技術涉及第一節點，其中所述硬件鏈路定時器被配置為基於與FIFO存儲器的條目相關聯的活動鏈路的數量調整時間段，並且其中活動鏈路包括第一鏈路和第二鏈路。

在一些方面中，本文中描述的技術涉及第一節點，其

中所述硬件鏈路定時器被配置為：基於存儲在FIFO存儲器的第一條目中的與第一鏈路相關聯的時序和狀態信息，確定重放第一多個分組中的至少一些；以及基於存儲在FIFO存儲器的第二條目中的與第二鏈路相關聯的時序和狀態信息，確定退出第二多個分組。

在一些方面中，本文中描述的技術涉及第一節點，其中所述第二多個分組存儲在第一節點的本地存儲裝置中，並且其中所述硬件鏈路定時器使本地存儲裝置響應於確定退出所述第二多個分組而丟棄所述第二多個分組。

在一些方面中，本文中描述的技術涉及第一節點，其中與第一鏈路相關聯的時序和狀態信息指示第一節點在用於重放分組的閾值持續時間內未接收到對接收第一多個分組之一的確認。

在一些方面中，本文中描述的技術涉及一種在基於乙太網的網絡中的第一節點處實現的計算機實現的方法，該計算機實現的方法包括：在第一節點的先進先出(FIFO)存儲器中存儲與多個鏈路相關聯的時序和狀態信息，其中所述第一節點被配置為使用乙太網協議在多個鏈路上向一個或多個其他節點傳輸分組；基於硬件定時器的相應滴答來訪問FIFO存儲器的條目；以及基於與多個鏈路中的第一鏈路相關聯的時序和狀態信息，確定重放與第一鏈路相關聯的至少一個分組，其中乙太網協議是有損的。

在一些方面中，本文中描述的技術涉及計算機實現的方法，其中以輪詢方式訪問FIFO存儲器的條目。

在一些方面中，本文中描述的技術涉及計算機實現的方法，進一步包括：基於與FIFO存儲器的條目相關聯的活動鏈路的數量，調整硬件定時器的時間段，其中活動鏈路被包括在多個鏈路中。

在一些方面中，本文中描述的技術涉及計算機實現的方法，進一步包括：基於與多個鏈路中的第二鏈路相關聯的時序和狀態信息，確定退出與第二鏈路相關聯的分組。

在一些方面中，本文中描述的技術涉及計算機實現的方法，進一步包括使與第一鏈路相關聯的至少一個分組被重放。

在一些方面中，本文中描述的技術涉及計算機實現的方法，其中與多個鏈路中的第一鏈路相關聯的時序和狀態信息指示第一節點未在用於重放分組的閾值持續時間內接收到對接收與第一鏈路相關聯的至少一個分組的確認。

在一些方面中，本文中描述的技術涉及上面描述和討論的所有實施例。

【圖式簡單說明】

貫穿圖式，重複使用圖式標記來指示所引用元素之間的對應性。提供圖式來說明本文中描述的主題的示例，並且不是限制其範圍。

參考隨附圖式描述了本公開的實施例，其中類似的圖式標記引用類似的元素，並且其中：

[圖1A]-[圖1B]是示出在開放系統互連(OSI)模型的不

同層上操作的示例協議的表格。

[圖 2]描繪了根據本公開的實施例的用於打開和關閉實現特斯拉傳輸協議(TTP)的節點之間的鏈路的示例狀態機。

[圖 3A]-[圖 3B]是描繪了根據本公開的實施例實現 TTP 的兩個設備之間的分組的傳輸和接收的示例時序圖。

[圖 4]圖示了根據本公開的實施例實現 TTP 的節點的示例示意性框圖。

[圖 5]描繪了根據本公開的實施例的依照 TTP 傳輸或接收的分組的示例報頭。

[圖 6]圖示了其中可以實現本公開的實施例的示例網絡和計算環境。

[圖 7A]-[圖 7B]示出了根據本公開的一些實施例的不同類型的 TTP 分組的操作碼。

[圖 8]圖示了根據本公開的一些實施例的用於存儲分組的示例物理存儲裝置，所述分組用於重放在有損協議(諸如 TTP)下傳輸和/或接收的分組。

[圖 9]描繪了根據本公開的一些實施例的用於跟蹤和維護用於傳輸和重放分組的傳輸次序的示例數據結構(例如，鏈表)。

[圖 10]圖示了根據本公開的一些實施例的用於重放在多個鏈路上傳輸的分組的硬件重放架構的至少一部分的示例框圖。

[圖 11]圖示了根據本公開的一些實施例的硬件鏈路定

時器的示例框圖，該硬件鏈路定時器實現超時檢查機制，用於在沒有軟件輔助的情況下重放分組。

[圖 12]圖示了根據本公開的一些實施例的用於重放從節點傳輸的分組的說明性例程。

[圖 13]描繪了用於確定是否重放與節點相關聯的一個或多個鏈路的示例例程。

【實施方式】

以下具體實施方式呈現了具體實施例的各種描述。然而，本文中描述的創新可以以多種不同的方式體現，例如，如申請專利範圍所限定和覆蓋的。在本說明書中，參考了圖式，其中類似的圖式標記和/或術語可以指示相同或功能相似的元素。將理解，圖中所圖示的元素不一定是按比例繪製的。此外，將理解，某些實施例可以包括比圖式中所圖示更多的元素和/或圖式中所圖示元素的子集。另外，一些實施例可以併入來自兩個或更多個圖式的特徵的任何合適的組合。標題僅為方便起見而提供，並且不影響申請專利範圍的範圍或含義。

一般而言，本公開的一個或多個方面對應於使用硬件機制(例如，無需軟件的輔助)控制網絡業務的系統和方法。更具體地，本公開的一些實施例公開了與乙太網標準兼容並且可通過硬件電路實現的流控制協議，以實現低時延，諸如一位數微秒內的時延。在一些實施例中，至少部分地通過利用硬件控制的狀態機來簡化網絡節點之間的通

信鏈路的打開和關閉，來實現一位數微秒時延。此外，所公開的流控制協議(例如，特斯拉傳輸協議(TTP))可以限制在轉換到硬件控制的狀態機的下一個狀態之前在已建立的鏈路上傳輸/重傳的分組的數量和/或等待時段的持續時間。這有助於實現低通信時延。有利的是，本文中公開的流控制協議實現了開放系統互連(OSI)模型的高達第四層(傳輸層)的純硬件實施方式。

本公開的一些方面涉及設計為在僅硬件上運行的流控制。這樣的流控制可以在沒有軟件流控制或中央處理單元(CPU)/內核參與的情況下實現。這可以允許IEEE 802.3乙太網能力，其時延僅受物理限制或主要受物理限制。例如，可以實現一位數微秒的時延。

特斯拉乙太網傳輸協議(TTP)是僅硬件乙太網流控制協議，其可以實現高達OSI模型中的傳送層。第2層(L2)乙太網流控制可以以僅硬件實現。第3層和/或第4層乙太網流控制也可以以僅硬件實現。鏈路控制、定時器、擁塞和重放功能可以在硬件中實現。TTP可以在網絡接口處理器和網絡接口卡中實現。TTP可以啟用完整的I/O批處理配置。TTP是一種有損協議。在有損協議中，丟失的數據可以恢復。例如，在有損協議中，任何丟失或損壞的分組都可以被重放(例如，重傳)和恢復，直到接收被確認。

本公開中的L2報頭、狀態機和操作碼可以定義該僅硬件協議(例如，TTP)，該僅硬件協議可以從N對N鏈路集中的丟失分組中恢復。

此外，本公開的一些實施例公開了一種硬件重放架構(例如，微架構)，其能夠重放在有損協議(諸如，**TTP**)下傳輸和/或接收的分組。如上所述，**TTP**(或**TTPoE**)是一種僅硬件乙太網流控制協議。**TTP**可以促進**HPC**和/或**AI**訓練系統的極低時延(例如，(一個或多個)一位數微秒)結構的實施方式。為了在沒有軟件控制機制的輔助下實現有損乙太網流控制協議，本公開的一些方面描述了一種硬件重放架構，其可以緩衝、保持、確認和/或重放分組，使得任何丟失或損壞的分組可以被重放和恢復，直到接收被確認為止。

為了利用僅硬件資源重放依照諸如**TTP**之類的有損乙太網協議傳輸的和/或接收的分組，所公開的硬件重放架構的一些實施例利用物理存儲裝置和數據結構存儲在不同鏈路中傳輸和/或接收的分組，並維護傳輸的分組的次序，特別是在重放發生時。在一些實施例中，物理存儲裝置可以是存儲、緩衝或保持與一個或多個鏈路相關聯的分組的任何類型的本地存儲裝置或高速緩存(例如，低級高速緩存)。物理存儲裝置的大小可能有限，諸如具有兆字節(**MB**)或千字節(**KB**)數量級的大小。在一些實施例中，數據結構可以包括一個或多個鏈表，其中每個鏈表可以記錄和/或跟蹤為在第一通信節點和第二通信節點之間建立的鏈路傳輸的分組的次序。有利的是，使用硬件重放架構實現用於有損協議的重放機制，該硬件重放架構採用大小有限的物理存儲裝置和跟蹤各種鏈路的分組次序的鏈表，

這允許通信節點在有限的硬件資源下(例如，當虛擬處理或存儲資源不可用時)遵照TTP進行操作。

另外，本公開的一些實施例涉及一種硬件鏈路定時器，其在沒有軟件控制機制的輔助的情況下實現超時檢查。本公開的一些方面描述了一種硬件鏈路定時器，其採用能夠通過與先進先出(FIFO)存儲器的協調來在多個鏈路上跟蹤超時的單個定時器，而不是採用多個定時器來在每個鏈路的基礎上跟蹤超時。更具體地，FIFO存儲器的條目可以存儲鏈路的狀態和/或定時器信息，並且硬件鏈路定時器可以以輪詢方式訪問FIFO存儲器的條目，以確定與鏈路相關聯的分組是可以被丟棄還是需要被保留。如果硬件鏈路定時器確定與該鏈路相關聯的分組可以被丟棄，則在受限的硬件資源下，更多的空間可以可用於存儲與另一鏈路相關聯的分組。如果硬件鏈路定時器確定應當保留與該鏈路相關聯的一個或多個分組，則與該鏈路相關聯的(一個或多個)保留的分組可以使得託管硬件鏈路定時器的通信節點能夠重放所述(一個或多個)保留的分組。

乙太網是有線通信的已建立標準技術。近年來，乙太網也已經發現在汽車工業中用於各種車輛應用。通常，與乙太網通信相關聯的時延範圍從數百微秒到超過若干毫秒。除了物理限制(例如，通信介質上的信號傳播速度)之外，用於控制乙太網上數據流的相關聯協議的複雜性通常已經呈現出另一個時延瓶頸。例如，為了遵循傳輸控制協議(TCP)或用戶數據報協議(UDP)，一般可能期望軟件控制

的管理。軟件控制或軟件輔助的網絡流控制管理往往增加與通信相關聯的時延。

然而，這樣的時延限制可能使乙太網技術不太適用於諸如高性能計算(HPC)和人工智能(AI)訓練數據中心之類的應用，在這些應用中，單微秒內的時延可能是改善系統性能和效率所期望的。儘管諸如融合乙太網(RoCE)上的遠程直接存儲器訪問(RDMA)或乙太網上的InfiniBand(IBoE)之類的協議可能有助於減少時延，但是它們可能帶來更大的系統設計複雜性或成本。例如，RoCE或InfiniBand具有無損網絡和擴展規範，其可能實現起來有挑戰性。實現RoCE或InfiniBand也可能導致顯著的軟件控制開銷或涉及帶寬受限的集中式令牌控制機制。此外，實現RoCE或InfiniBand的系統可能暫停頻繁(例如，頻繁暫停)。

為解決上述問題中的至少一部分，本公開的一些實施例公開了可在基於乙太網的網絡或點對點(P2P)網絡上操作的流控制協議(例如，特斯拉傳輸協議(TTP))。流控制協議可以完全通過硬件可實現，而沒有軟件控制機制的輔助，以便將通信的時延帶到一位數微秒內。流控制協議可以在不涉及軟件資源的情況下實現，諸如執行計算機可讀指令或操作系統的通用處理器或中央處理單元。此外，在一些機制(例如，在暫停之前限制可以傳輸的分組的數量、限制可以同時建立的鏈路的數量、硬件控制的狀態機、或依照TTP傳輸或接收的分組的建議報頭格式中的一個或多個)構建到流控制協議中的情況下，不需要虛擬化

資源(例如，虛擬化處理器或存儲器)來實現流控制協議。

在一些實施例中，狀態機加速不同狀態之間的轉換，以打開和關閉節點之間的通信鏈路。狀態機可以由硬件維護和實現，而不涉及軟件、固件、驅動程序或其他類型的可編程指令。照此，與利用軟件支持的其他協議(諸如可適用於基於乙太網的網絡的傳輸控制協議(TCP))的實施方式相比，可以加速狀態機的不同狀態之間的轉換。

在一些實施例中，依照TTP傳輸和接收的分組的報頭(例如，TTP報頭)支持開放系統互連(OSI)模型的從第2層至第4層的操作。報頭可以包括由現有的基於乙太網的網絡設備或基礎設施可識別的字段。照此，可以保留TTP與現有乙太網標準的兼容性。有利地，這可以允許現有基礎設施和/或供應鏈的經濟使用，帶來更多的系統設計選項，並實現系統級的重用或冗餘。

如上所述，節點可以使用僅硬件資源在TTP下實現或操作(例如，使用TTP與另一節點通信)，而無需軟件控制機制的輔助。為了在具有純硬件資源的TTP下操作，節點可以採用硬件重放架構來重放可能在傳輸中丟失的分組。在一些實施例中，硬件重放架構可以包括本地存儲裝置，諸如用於存儲在一個或多個鏈路上傳輸和/或接收的分組的一個或多個高速緩存，其中一個或多個鏈路中的每一個可以依照TTP打開或關閉。與諸如TCP或UDP之類的協議——在這些協議中，具有幾乎無限的處理功率和存儲容量的虛擬化資源通常可通過軟件控制的網絡流控制管理獲得

——相反，在 **TTP** 下操作的節點內的硬件重放架構所採用的高速緩存(例如，低級高速緩存)的大小可能是有限的。例如，高速緩存的大小可以是兆字節(**MB**)或千字節(**KB**)的數量級，諸如 **256 KB**。為了在有限的本地存儲裝置下通過依照諸如 **TTP** 之類的有損通信協議建立的一個或多個鏈路彼此通信，與一個或多個鏈路相關聯的分組應該被充分管理(例如，保留或丟棄)，使得一些分組被保留用於重放，而其他分組被丟棄以避免高速緩存溢出。

在一些示例中，使用 **TTP** 下建立的鏈路向第二節點傳輸 **N** 個分組的第一節點可以利用高速緩存存儲 **N** 個分組，**N** 為可以受高速緩存大小限制的任意正整數。只要來自 **TTP** 和/或網絡條件的約束准許，第一節點可以連續地向第二節點傳輸 **N** 個分組中的一些或全部。為了容納重放的分組，高速緩存可以繼續存儲已經傳輸的分組，直到從第二節點接收到對接收到分組的確認。當接收到對接收到分組的確認時，高速緩存可以丟棄該分組，以騰出空間來存儲要在第一節點和第二節點或其他節點之間的鏈路或其他鏈路上傳輸的分組。相反，如果接收到對分組的未確認(例如，第二節點通知第一節點沒有接收到分組)或者在沒有從第二節點接收到對接收到分組的確認或未確認的情況下發生超時，則第一節點可以重放分組(例如，向第二節點重傳分組)。與重放分組相關聯，第一節點可以丟棄已經接收到對接收的確認的其他分組。

在一些示例中，傳輸和重放分組的次序可以相同。例

如，第一節點可以以特定次序傳輸 N 個分組(例如，第1個分組、第2個分組到第 N 個分組)。如果重放第5個分組(例如，響應於第一節點從第二節點接收到對第5個分組的未確認，響應於在沒有接收到確認或沒有對接收到第5個分組的確認的情況下發生超時)並且已經接收到關於第1到第4個分組的確認，則高速緩存可以丟棄第1到第4個分組而不是第5個分組，使得節點可以重放第5個分組。附加地和/或可選地，當重放第5個分組時，第一節點可以以與先前傳輸的次序相同的次序重放在第5個分組之後傳輸的分組(假設 $N>5$)。

在一些示例中，第一節點的硬件重放架構可以利用與高速緩存協調的鏈表，以維護 N 個分組中的一些或全部的首次傳輸與之後的任何重放之間的次序。鏈表可以包括 N 個元素，其中每個元素包括 N 個分組中的每一個以及對對應於下一個分組的下一個元素的引用。當傳輸和/或重放 N 個分組時，硬件重放架構可以進一步利用指向鏈表中的一個或多個元素的一個或多個指針來確定分組是要被保留用於重放還是可以被丟棄(例如，為了節省存儲資源)。取 N 為9(例如，從第一節點傳輸到第二節點的9個分組)作為示例，在鏈表中，第1個元素可以包括第1個分組和第1個引用，其中第1個引用指向第2個元素；第2個元素可以包括第2個分組和第2個引用，其中第2個引用指向第3個元素；並且第8個元素可以包括第8個分組和第8個引用，其中第8個引用指向第9個元素；並且第9個元素可以包括第9個分

組。硬件重放架構可以維護和更新指向三個元素的三個指針。假設該節點已經傳輸了第1到第9個分組，並且已經從第二節點接收到對接收到第1到第7個分組而不是第8和第9個分組的確認，則第一指針可以指向鏈表的第1個元素，第二指針可以指向鏈表的第8個元素，並且第三指針可以指向鏈表的第9個元素。照此，硬件重放架構可以使高速緩存基於這三個指針丟棄分組並重放分組。更具體地，高速緩存可以通過第三指針所指向的分組(例如，第9個分組)來重放第二指針所指向的分組(例如，第8個分組)，並且丟棄剩餘的分組(例如，第二指針所指向的分組之前的第一指針所指向的分組)。附加地和可選地，一些或所有硬件重放架構可以以流水線方式操作，以增加節點的吞吐量。有利的是，使用高速緩存和鏈表來實現重放功能使得第一節點能夠在有限的硬件資源下使用TTP與第二節點通信，而無需軟件控制機制的輔助。

如上所述，在TTP協議下操作的節點可以包括硬件鏈路定時器，以實現超時檢查機制，用於在沒有軟件輔助的情況下重放分組。與其它乙太網協議(例如，TCP或UDP)相反——軟件通常採用所述其它乙太網協議使用多個定時器(例如，一個定時器用於一個鏈路)來跟蹤多個鏈路上的超時——硬件鏈路定時器可以允許節點確定哪個(哪些)分組在哪個(哪些)鏈路上傳輸以進行重放，並且如果期望重放，則在有限的硬件資源下何時進行重放(例如，當虛擬和/或物理地址空間和計算資源的大型資源池不可用時)。

在一些實施例中，硬件鏈路定時器可以週期性地對與節點相關聯的已建立鏈路(例如，活動鏈路)執行時序檢查。硬件鏈路定時器可以包括先進先出(FIFO)存儲器，該先進先出存儲器可以存儲與每個活動鏈路相關聯的時序和狀態信息，並且以輪詢方式檢查與每個活動鏈路相關聯的時序和狀態。硬件鏈路定時器可以利用單個可編程定時器來為多個活動鏈路和/或分組調度時間點，以讀出與多個活動鏈路和/或分組中的每一個相關聯的時序和狀態信息。讀出的時序和狀態信息可以用於通過進一步的信息查找來確定是重放與鏈路相關聯的分組還是丟棄分組。

在一些示例中，FIFO存儲器可以存儲與第一節點和(一個或多個)其他節點之間建立的一個或多個鏈路相關聯的時序信息。例如，第一節點可以包括硬件鏈路定時器，其使用FIFO存儲器來存儲與在第一節點和一個或多個其他節點之間建立的M個鏈路相關聯的時序信息，其中M是大於一的正整數。代替使用M個定時器，其中每個定時器跟蹤對應鏈路的時序信息，硬件鏈路定時器可以利用單個定時器(例如，在可編程的時間段內滴答一次的定時器)通過以輪詢(例如，循環)方式訪問FIFO存儲器來跟蹤和/或更新M個鏈路中的每一個的時序信息。具體地，當單個定時器滴答一次時，硬件鏈路定時器可以一次一個地訪問FIFO存儲器的條目，其中FIFO存儲器的每個被訪問的條目對應於M個鏈路之一。在一些實施例中，每個滴答的時間段可以變化，並且可以在數百微秒到一位數微秒的數量級。例

如，一個滴答的時間段可以高達100微秒，並且可以低至1微秒。此外，硬件鏈路定時器可以基於由FIFO存儲器的條目表示的鏈路數量(例如， M)來調整滴答的時間段。例如，當 M 增加(例如，FIFO存儲器的條目表示更多的鏈路)時，滴答的時間段可能減少；並且當 M 減小(例如，FIFO存儲器的條目表示更少的鏈路)時，滴答的時間段可能增加。照此，如果滴答的時間段與FIFO存儲器的條目所表示的鏈路數量不成比例地改變，則在其內檢查鏈路的狀態和/或時序信息的時間間隔可以保持不變。

在一些示例中，與 M 個鏈路之一相關聯的時序和/或狀態信息可以指示該鏈路已經有多長時間未接收到對被傳輸的接收到分組的確認。假設第一節點已經在鏈路上向第二節點傳輸了 N 個分組，FIFO存儲器的一個條目可以存儲時序和/或狀態信息，該時序和/或狀態信息當在一個滴答的特定時間段下通過輪詢方式被訪問時指示在預確定的持續時間內沒有接收到對接收到 N 個分組中的任何一個的確認。在訪問FIFO存儲器的條目時，硬件鏈路定時器可以利用存儲在條目中的時序和/或狀態信息來查找可以存儲在第一節點的本地存儲裝置(例如，低級高速緩存)中的 N 個分組，以用於重放這 N 個分組。替代地，與 M 個鏈路之一相關聯的時序和/或狀態信息可以存儲在FIFO存儲器的一個條目中，以指示該鏈路可以關閉(例如，第一節點傳輸的所有分組都已經被第二節點接收)。在訪問FIFO存儲器的條目時，硬件鏈路定時器可以利用存儲在條目中的時序

和/或狀態信息來查找可能仍然存儲在第一節點的本地存儲裝置中的分組，並且丟棄該分組，因為存儲在FIFO存儲器的條目中的時序和/或狀態信息指示可以關閉鏈路。有利的是，通過針對多個鏈路和/或在可調整週期下滴答的分組利用單個定時器以及存儲多個鏈路的時序和/或狀態信息的FIFO存儲器，第一節點可以以適當的時序重放分組，以實現低時延並釋放由非活動鏈路(例如，關閉的鏈路)佔用的硬件資源，以供活動鏈路使用，從而在有限的計算和存儲資源下操作。

儘管將根據說明性實施例和特徵組合對各個方面進行描述，但是相關領域的技術人員將領會，示例和特徵組合本質上是說明性的，並且不應解釋為限制性的。更具體地，本申請的各方面可以適用於不同背景下的各種類型的網絡和通信協議。更進一步，儘管將描述用於控制網絡流的電路框圖或狀態機的具體架構，但是這樣的說明性的電路框圖或狀態機或架構不應被解釋為限制性的。因此，相關技術領域的技術人員將領會，本申請的各方面不一定限於應用於任何特定類型的網絡、網絡基礎設施、或網絡節點之間的說明性交互。

特斯拉傳輸協議(TTP)

圖1A-圖1B是示出OSI模型(具有七層)連同與每層相關聯的示例協議的表格。圖1A示出了在OSI模型的第4層(例如，傳輸層)上操作的TCP和UDP協議的示例協議。圖1B示

出了在 OSI 模型的第 4 層上操作的特斯拉傳輸協議 (TTP) 的示例協議。

如圖 1A 中所示，除了在第 4 層上操作的 TCP 或 UDP，連同 TCP 或 UDP 一起操作的其他示例協議或應用可以包括：在第 7 層上操作的超文本傳送協議 (HTTP)、電傳網絡 (Telnet)、文件傳送協議 (FTP)；在第 6 層上操作的聯合圖像專家組 (JPEG)、便攜式網絡圖形 (PNG)、移動圖像專家組 (MPEG)；在第 5 層上操作的網絡文件系統 (NFS) 和結構化查詢語言 (SQL)；在第 3 層上操作的互聯網協議版本 4 (IPv4)/互聯網協議版本 6 (IPv6)；等等。對於 TCP 或 UDP 在第 4 層上操作，第 4 層的實施方式通常涉及如圖 1A 中所示的軟件。

如圖 1B 中所示，除了在第 4 層上操作的 TTP，連同 TTP 一起操作的其他示例協議或應用可以包括：在第 7 層上操作的 Pytorch；在第 6 層上操作的 FFMPEG、高效視頻編碼 (HEVC)、YUV；在第 5 層上操作的 RDMA；在第 3 層上操作的 IPv4/IPv6；等等。與圖 1A 形成對照，對於 TTP 在第 4 層上操作，OSI 模型的第 1 層到第 4 層的實施方式可以以僅硬件實行，而不涉及如圖 1B 中所示的軟件。有利的是，與圖 1A 中所示的實施方式相比，通過如圖 1B 中所示的基於 TTP 的 OSI 模型的第 1 層至第 4 層的純硬件實施方式可以縮短基於乙太網的網絡上的通信的時延。

圖 2 描繪了根據本公開的實施例的用於打開和關閉實現 TTP 的節點之間的鏈路的示例狀態機 200。狀態機 200 可

以由網絡接口處理器或網絡接口卡來實現。對於在乙太網鏈路上通信的每個節點上的節點之間的每個乙太網鏈路，可以存在一個狀態機 200。例如，如果網絡接口處理器可以在 5 個 TTP 鏈路上與 5 個網絡接口卡通信，則網絡接口處理器可以包括狀態機 200 的 5 個實例，其中每個鏈路一個實例。在這個示例中，5 個網絡接口卡中的每一個可以具有狀態機 200 的一個實例，用於與網絡接口處理器通信。在一些實施例中，使用狀態機 200 彼此通信的節點可以形成對等網絡。

如圖 2 中所示，狀態機 200 包括關閉狀態 202、打開接收狀態 204、打開發送狀態 206、打開狀態 208、關閉接收狀態 210 和關閉發送狀態 212。狀態機 200 可以在關閉狀態 202 下開始，這可以指示在維護狀態機 200 的第一節點和要與之建立通信鏈路的第二節點之間當前沒有通信鏈路打開。另外，狀態機 200 的單獨副本可以由基於本公開中公開的特斯拉傳輸協議 (TTP) 操作的節點來維護、更新和轉換。此外，如果基於 TTP 操作的節點與多個節點同時通信或在時間上重疊，則該節點可以為每個鏈路保留多個且獨立的狀態機 200。

狀態機 200 然後可以取決於第一節點是否向第二節點傳輸或從第二節點接收建立通信鏈路的請求而不同地轉換。如果第一節點向第二節點傳輸打開通信鏈路的請求，則狀態機 200 可以從關閉狀態 202 轉換到打開發送狀態 206。另一方面，如果第一節點從第二節點接收到打開通

信鏈路的請求，則狀態機200可以從關閉狀態202轉換到打開接收狀態204。

當處於打開發送狀態206時，狀態機200可以停留在打開發送狀態206，或取決於各種準則轉換回到關閉狀態202或前進到打開狀態208。如果第一節點從第二節點接收到open-nack(例如，拒絕打開鏈路的請求的消息)，則狀態機200可以從打開發送狀態206轉換回到關閉狀態202。另一方面，如果第一節點從第二節點接收到open-ack(接受打開鏈路的請求的消息)，則狀態機200可以從打開發送狀態206轉換到打開狀態208。替代地，如果第一節點在特定時間段內沒有從第二節點接收到open-nack或open-ack，則第一節點可以超時，然後第一節點可以向第二節點重傳打開通信鏈路的請求，並停留在打開發送狀態206。

如上所述，當處於關閉狀態202時，如果第一節點從第二節點接收到打開通信鏈路的請求，則狀態機200可以從關閉狀態202轉換到打開接收狀態204。在打開接收狀態204下，狀態機200可以取決於第一節點是接受還是拒絕來自第二節點的打開鏈路的請求而不同地轉換。例如，第一節點可以選擇向第二節點傳輸open-nack(例如，拒絕打開鏈路的請求)。在這樣的情況下，狀態機200可以轉換回到關閉狀態202，其中第一節點可以進一步傳輸或接收來自第二節點或其他節點的打開鏈路的請求。替代地，在打開接收狀態204下，第一節點可以向第二節點傳輸open-ack，並且然後轉換到打開狀態208。

當處於打開狀態 208 時，第一節點和第二節點可以通過建立的通信鏈路彼此傳輸和接收分組。該鏈路可以是有線乙太網鏈路。第一節點可以停留在打開狀態 208 下，直到一些條件發生為止。在一些實施例中，狀態機 200 可以響應於接收到關閉允許第一節點和第二節點在打開狀態 208 下時傳輸和接收分組的通信鏈路的請求，從打開狀態 208 轉換到關閉接收狀態 210。替代地，響應於第一節點向第二節點傳輸關閉通信鏈路的請求，狀態機 200 可以從打開狀態 208 轉換到關閉發送狀態 212。除了請求關閉通信鏈路之外，如果通信鏈路已經空閒超過閾值時間量，則狀態機 200 可以從打開狀態 208 轉換到關閉接收狀態 210 或關閉發送狀態 212。

當處於關閉接收狀態 210 時，如果第一節點向第二節點傳輸 `close-ack` (例如，確認或接受關閉鏈路請求的消息)，則狀態機 200 可以轉換回到關閉狀態 202。否則，如果第一節點向第二節點傳輸 `close-nack` (例如，拒絕或不確認關閉鏈路的請求的消息)，則狀態機 200 可以停留在關閉接收狀態 210。

當處於關閉發送狀態 212 時，如果第一節點從第二節點接收到 `close-ack` (例如，確認或接受關閉鏈路請求的消息)，則狀態機 200 可以轉換回到關閉狀態 202。否則，如果第一節點接收到從第二節點傳輸的 `close-nack` (例如，拒絕或不確認關閉鏈路的請求的消息)，則狀態機 200 可以停留在關閉發送狀態 212。在關閉發送狀態 212 下，如果第一

節點在超時閾值內沒有收到來自第二節點的回應，則第一節點可以向第二節點重新發送關閉通信鏈路的請求。

在一些實施例中，狀態機200可以由硬件維護和實現，而不涉及軟件、固件、驅動程序、或其他類型的可編程指令。照此，與涉及軟件支持的其他協議(諸如適用於基於以太網的網絡的傳輸控制協議(TCP))的實施方式相比，狀態機200的不同狀態之間的轉換可以被加速。

在一些實施例中，第一節點可以立即停止傳輸傳輸隊列中的分組，並在處於關閉接收狀態210時，響應於從第二節點接收到關閉鏈路的請求，向第二節點發送close-ack，而不是保持傳輸分組等待被傳輸並存儲在傳輸隊列中。有利的是，在接收到關閉鏈路的請求之後，在無限期的時間量內避免繼續傳輸分組使得第一節點能夠從打開狀態208轉換回到關閉狀態202，具有更少的轉換週期和更少的時間不確定性。

此外，在打開狀態208期間，可以由第一節點或第二節點連續傳輸的分組的數量可以受到限制。例如，當處於打開狀態208時，第一節點在停止傳輸分組之前可以僅連續傳輸N個分組，其中N可以是從1到超過一千的正整數。數字N可以被物理存儲器所限制。在一些實施例中，N可能受到第一節點可用的物理存儲器(例如，動態隨機存取存儲器等)的大小限制或約束。具體地，N可以與關聯於第一節點或第二節點的物理存儲器的大小成比例。例如，如果1千兆字節(GB)的物理存儲器被分配給第一節點，則N

可能高達一百萬。在一些實施例中， N 可以在數萬或數十萬之內。在打開狀態208期間，可以跟蹤用於交換分組的物理存儲器的量。有利的是，限制可以由第一節點或第二節點連續傳輸的分組的數量可以減少實現狀態機200的計算和存儲資源。與一般通過虛擬化(例如，虛擬化的存儲器或處理資源)假定無限的軟件和硬件資源的可用性的協議(例如，TCP)形成對照，限制傳輸的分組的數量允許TTP在更受約束的計算和存儲資源下操作。

在一些實施例中，第一節點或第二節點在接收到close-ack或向另一方傳輸close-ack之後，不進一步等待關閉鏈路。例如，當處於關閉發送狀態212時，響應於接收到從第二節點傳輸的close-ack，第一節點可以立即轉換到關閉狀態202。第一節點可以在更短的時間量內從關閉發送狀態212轉換回到關閉狀態202，而不是等待另一個預確定或隨機的時間段來監控第二節點是否有附加的分組要被傳輸。有利的是，這增加了精度並縮短了與狀態機200的狀態之間的轉換相關聯的時延，從而允許TTP以低於諸如TCP之類的協議的時延來促進通信。

圖3A-圖3B圖示了示例時序圖，其描繪了實現根據本公開的實施例的TTP的兩個設備之間的分組的傳輸和接收。圖3A圖示了從設備A向設備B傳輸的分組沒有丟失的一種場景，而圖3B圖示了從設備A向設備B傳輸的一些分組丟失的另一種場景。圖3A-圖3B可以結合狀態機200來理解。設備A和設備B是在TTP上通信的兩個示例節點。

如圖 3A 中所示，處於關閉狀態 202 的設備 A 可以向設備 B 傳輸分組 ID=0 的 TTP_OPEN。在 (1) 處向設備 B 傳輸 TTP_OPEN 之後，由設備 A 維護的狀態機可以從關閉狀態 202 轉換至打開發送狀態 206。此外，在 (1) 處從設備 A 接收到 TTP_OPEN 之後，由設備 B 維護的狀態機可以從關閉狀態 202 轉換到打開接收狀態 204。

然後，在 (2) 處從設備 B 接收到 TTP_OPEN_ACK 之後，由設備 A 維護的狀態機可以從打開發送狀態 206 轉換到打開狀態 208。此外，在 (2) 處將 TTP_OPEN_ACK 傳輸到設備 A 之後，由設備 B 維護的狀態機可以從打開接收狀態 204 轉換到打開狀態 208。

在 (3) 處，當處於打開狀態 208 時，在從設備 B 接收到任何響應之前，設備 A 可以向設備 B 不斷地或連續地傳輸四個分組 (例如，TTP_PAYLOAD ID=1 至 4)。在一些實施例中，在從設備 B 接收到任何響應之前，設備 A 可以向設備 B 傳輸的分組的數量是有限的。響應於在 (4) 處從設備 A 接收到的分組，設備 B 可以傳輸四個分組 (例如，TTP_ACK ID=1 至 4)，從而確認接收到由設備 A 傳輸的四個分組。

在 (5) 處，設備 A 將 TTP_CLOSE (其中分組 ID = 5) 傳輸至設備 B。傳輸 TTP_CLOSE 之後，由設備 A 維護的狀態機可以從打開狀態 208 轉換至關閉發送狀態 212。響應於從設備 A 接收到 TTP_CLOSE，由設備 B 維護的狀態機可以從打開狀態 208 轉換到關閉接收狀態 210。

此後，在 (6) 處，設備 B 可以向設備 A 傳輸

TTP_CLOSE_ACK(其中分組ID=5)。在向設備A傳輸TTP_CLOSE_ACK之後，由設備B維護的狀態機可以從關閉接收狀態210轉換回到關閉狀態202。在從設備B接收到TTP_CLOSE_ACK之後，由設備A維護的狀態機可以從關閉發送狀態212轉換回到關閉狀態202。照此，設備A和設備B之間的鏈路/連接可能是關閉的。

圖3B圖示了與本公開中公開的流控制協議(例如，TTP)相關聯的“有損”流控制特徵，其中有損可以指示在接收到未確認之後重傳丟失或損壞的分組。

如圖3B中所示，處於關閉狀態202時的設備A可以向設備B傳輸分組ID=0的TTP_OPEN。在(1)處向設備B傳輸TTP_OPEN之後，由設備A維護的狀態機可以從關閉狀態202轉換至打開發送狀態206。此外，在(1)處從設備A接收到TTP_OPEN之後，由設備B維護的狀態機可以從關閉狀態202轉換到打開接收狀態204。

然後，在(2)處從設備B接收到TTP_OPEN_ACK之後，由設備A維護的狀態機可以從打開發送狀態206轉換到打開狀態208。此外，在(2)處將TTP_OPEN_ACK傳輸到設備A之後，由設備B維護的狀態機可以從打開接收狀態204轉換到打開狀態208。

在(3)處，當處於打開狀態208時，在從設備B接收到任何響應之前，設備A可以不斷地或連續地向設備B傳輸四個分組(例如，TTP_PAYLOAD ID=1至4)。然而，由於一些網絡條件，設備B可能無法接收一些分組(例如，

TTP_PAYLOAD ID=3)。照此，在(4)處，設備B可以傳輸三個分組(例如，TTP_ACK ID=1到2，以及TTP_NACK ID=3)，從而確認接收到由設備A傳輸的兩個分組(ID=1到2)，但是通知沒有接收到具有TTP_PAYLOAD ID=3的分組。

在從設備B接收到分組(例如，TTP_NACK ID=3)之後，在(5)處，設備A向設備B重傳兩個分組(例如，TTP_PAYLOAD ID=3至4)。值得注意的是，在接收到分組(例如，TTP_NACK ID=3)之後，兩個分組的重傳反映了TTP的“有損”特徵。在一些實施例中，設備A可以在超時發生之後(例如，當本地計數器超過特定值時)重傳一些分組。有利的是，由於設備A和設備B之間的對等鏈接的存在，“有損”特徵使得TTP能夠無限制地控制或縮放網絡流，並且使得TTP能夠在預期丟失一些業務的大型系統中實現特定於鏈路的恢復。

在(6)處，在接收到兩個分組(例如，TTP_PAYLOAD ID=3至4)之後，設備B可以向設備A傳輸兩個分組(例如，TTP_ACK ID=3至4)，以確認接收到重傳的分組(例如，TTP_PAYLOAD ID=3至4)。

在(7)處，設備A可以向設備B傳輸分組(例如，TTP_CLOSE ID=5)，以嘗試關閉設備A和設備B之間的鏈路。此外，在(7)處，由設備A維護的狀態機可以從打開狀態208轉換至關閉發送狀態212，並且由設備B維護的狀態機可以從打開狀態208轉換至關閉接收狀態210。

在(8)處，設備B可以向設備A傳輸分組(例如，TTP_CLOSE_ACK ID=5)，以確認並同意關閉鏈路。由設備B維護的狀態機可以從關閉接收狀態210轉換回到關閉狀態202。響應於從設備B接收到分組(例如，TTP_CLOSE_ACK ID=5)，由設備A維護的狀態機可以從關閉發送狀態212轉換回到關閉狀態202。

在一些實施例中，設備A和/或設備B可能不轉換到打開狀態208，或者可能不傳輸或接收數據分組，直到協商鏈路的過程完成為止。例如，在設備A從設備B接收到TTP_OPEN_ACK之前，設備A可能不向設備B傳輸數據分組或從設備B接受數據分組。在這些實施例中，當關閉設備A和設備B之間的鏈路時，特別是當在設備A和設備B之間的先前鏈路關閉之後立即從設備A或設備B傳輸TTP_OPEN時，可能不存在強加超時時段的需要。

圖4圖示了根據本公開的實施例實現TTP的節點400的示例框圖。如圖4中所示，節點400可以包括發射(TX)路徑和接收(RX)路徑。如圖4中所示，在節點400的前端包括物理編碼子層(PCS)+物理介質附接(PMA)塊402，其處理OSI模型的第1層(例如，物理層)上的通信。在一些實施例中，PCS+PMA塊402基於頻率為156.25 MHz的參考時鐘404進行操作。在其他實施例中，PCS+PMA塊402可以在不同的時鐘頻率下操作。PCS+PMA塊402可以與乙太網或IEEE 802.3標準兼容。在用於處理RX路徑上的數據的操作中，PCS+PMA塊402接收RX serdes [3:0]作為輸入，並將

RX serdes [3:0]重新佈置成輸出(例如，RX幀408)，以由TTP媒體訪問控制(MAC)塊410處理。在用於處理TX路徑上的數據的操作中，PCS+PMA塊402從TTP MAC塊410接收TX幀412作為輸入，並重新佈置數據格式以輸出TX serdes [3:0]。

在RX路徑上，TTP MAC塊410接收RX幀408作為輸入，並向片上系統(SoC)420輸出RDMA接收數據416。在TX路徑上，TTP MAC塊410從SoC 420接收RDMA發送數據418，並將TX幀412輸出到PCS+PMA塊402。如圖4中所示，TTP MAC塊410可以處理OSI模型的第2層到第4層上的操作。TTP MAC塊410可以包括TTP有限狀態機(FSM)422。TTP FSM 422可以維護和更新如圖2中所示的狀態機200。如上面討論的，對於節點400與一個或多個其他節點建立的每個通信鏈路，TTP FSM 422可以維護和更新對應的狀態機(例如，狀態機200)以控制與相應通信鏈路相關聯的流。

在一些實施例中，PCS+PMA塊402和TTP MAC塊410可以由硬件實現，諸如以專用集成電路(ASIC)或現場可編程門陣列(FPGA)的形式。照此，PCS+PMA塊402和TTP MAC塊410可以在沒有軟件/固件/驅動程序的輔助或參與的情況下操作。有利的是，PCS+PMA塊402和TTP MAC塊410可以處理從OSI模型的第1層到第4層的通信，而無需軟件輔助，以減少與第1層到第4層中的通信相關聯的時延。

圖5描繪了依照TTP傳輸或接收的分組的示例報頭

500。如圖5中所圖示，示例報頭500具有64個字節。前16個字節包括用於乙太網第2層(例如，數據鏈路層)和虛擬局域網(VLAN)操作的報頭。第二個16字節包括ETHTYPE，隨後是可選的第3層互聯網協議(IP)報頭。為了支持基於TTP的第2層操作，可以將ETHTYPE設置為特定值(例如0x9AC6)。當ETHTYPE被設置為特定值時，報頭500可以向處理報頭500的網絡設備發信號通知報頭500是基於TTP格式化的。第三個16字節包括UDP下第3層(IP)操作和第4層操作的可選字段。在第三個16字節和第四個16字節的末尾是在TTP下第4層操作的字段。TTP可以被稱為乙太網上的TTP(TTPoE)。TTP在圖5中被標記為TTPoE。

有利的是，示例報頭500允許TTP支持OSI模型的從至少第2層至第4層的基於乙太網的網絡上的操作。具體地，現有的乙太網交換機和硬件可以支持與TTP相關聯的操作。

圖6圖示了其中可以實現本公開的實施例的示例網絡和計算環境600。示例網絡和計算環境600可以用於高性能計算或人工智能訓練數據中心。作為一個示例，網絡和計算環境600可以用於神經網絡訓練，以生成供車輛(例如，汽車)的自主駕駛系統使用的數據。如圖6中所示，示例網絡和計算環境600包括乙太網交換機608、主機602A到602E、外圍組件快速互連(PCIe)主機604A到604N、以及計算圖塊606A到606N。儘管在圖6中存在五個主機602A到

602E，但是可以實現多於或少於五個的任何合適數量的主機。此外，PCIe主機的數量和計算圖塊的數量可以是任何合適的正整數。

主機602A至602E中的每一個包括網絡接口卡(NIC)、中央處理單元(CPU)和動態隨機存取存儲器(DRAM)。儘管圖示為CPU，但是在一些實施例中，CPU可以體現為任何類型的單核、單線程、多核或多線程處理器、微處理器、數字信號處理器(DSP)、微控制器、或其他處理器或處理/控制電路。儘管圖示為DRAM，但是在一些實施例中，DRAM可以替代地或附加地體現為任何類型的易失性或非易失性存儲器或數據存儲裝置，諸如靜態隨機存取存儲器(SRAM)、同步DRAM(SDRAM)、雙倍數據速率同步動態隨機存取存儲器(DDR SDRAM)。DRAM可以存儲在主機602A至602E的操作期間使用的各種數據和程序代碼，包括操作系統、應用程序、庫、驅動程序以及諸如此類。

在一些實施例中，NIC可以實現TTP，用於與乙太網交換機608通信。每個NIC可以使用TTP作為流控制協議與乙太網交換機608通信，以管理經由乙太網交換機608在每個NIC和網絡接口處理器(NIP)之間建立的鏈路。在一些實施例中，NIC可以包括圖4的PCS+PMA塊402和TTP MAC塊410。在一些實施例中，NIC可以在沒有軟件/固件的輔助的情況下實現TTP。

如圖6中所示，PCIe主機604A至604N中的每一個可以包括網絡接口處理器(NIP)和高帶寬存儲器(HBM)。在一些

實施例中，由HBM支持的帶寬可以是每計算32千兆字節(GB)。PCIe主機604A至604N中的每一個可以與計算圖塊606A至606N中的每一個通信。計算圖塊606A至606N中的每一個可以包括存儲、輸入/輸出和計算資源。計算圖塊606A可以包括具有用於高性能計算的處理器陣列的晶片上系統。在一些應用中，計算圖塊606A至606N中的每一個可以執行每秒9 peta浮點運算(PFLOPS)，使用靜態隨機存取存儲器(SRAM)存儲大小為11千兆字節(GB)的數據，或者以每秒36太字節(TB)的帶寬促進輸入/輸出操作。

在一些實施例中，主機602A至602E中的每個NIC可以打開和關閉與PCIe主機604A至604N中的每個NIP的通信鏈路。具體地，一個NIC和一個NIP可以通過實現圖2的狀態機200來打開和關閉彼此之間的通信鏈路。為了打開和關閉通信鏈路，NIC和NIP可以使用包括圖7A-圖7B的操作碼的分組來執行期望的操作。例如，為了打開與NIP的鏈路，NIC可以向NIP傳輸包括操作碼TTP_OPEN(圖7A中所示)的分組，以請求打開通信鏈路。在接收到具有操作碼TTP_OPEN的分組之後，NIP可以從圖2的關閉狀態202轉換到打開接收狀態204。在發送具有操作碼TTP_OPEN_ACK(圖7A中所示)的分組之後，NIP可以從打開接收狀態204轉換到打開狀態208，如圖2中所圖示。在一些實施例中，一旦建立了通信鏈路(例如，當NIC和NIP都處於打開狀態208時)，NIC和NIP可以使用圖5的報頭500彼此傳輸或接收分組。換句話說，在NIC和NIP之間傳輸

或接收的每個分組可以包括圖5的報頭500。

如圖6中所指示，主機602A至602E中的每一個、PCIe主機604A至604N中的每一個、計算圖塊606A至606N中的每一個、或乙太網交換機608之間的通信和數據交換可以基於TTP進行。利用使用上述技術通過TTP實現的更短的時延(與TCP相比)，可以實現圖6的各種元素之間的高帶寬和高速通信。在一些實施例中，圖6中所示的至少一部分NIP或至少一部分NIC可以與圖4的節點400類似或相同地實現。儘管貫穿圖6沒有圖示，但是在一些實施例中，NIC和NIP中的每一個可以包括端口610，通過該端口610可以接收和傳輸分組。在一些實施例中，端口610是乙太網端口。

圖7A-圖7B示出了根據本公開的實施例的不同類型的TTP分組的操作碼。圖7A和圖7B中所示的TTP分組在圖2、圖3A和圖3B中用於關閉和打開網絡節點之間的鏈路。TTP分組可以在圖6的網絡和計算環境中的節點之間交換。結合圖2、圖3A和圖3B可以更好地理解圖7A和圖7B中所示的TTP分組。

分組重放硬件架構

回過來參考圖4，圖4圖示了使用TTP傳輸和/或接收分組的節點400的示例框圖，將描述重放硬件架構。如上所述，節點400可以包括諸如物理編碼子層(PCS)+物理介質附接(PMA)塊402和TTP媒體訪問控制(MAC)塊410之類的

塊，該塊包括TTP FSM 422，用於處理OSI模型的從第1層到第4層的通信，而無需軟件輔助來減少與第1層至第4層中的通信相關聯的時延。此外，節點400的TTP媒體訪問控制(MAC)塊410可以包括硬件重放架構，該硬件重放架構至少包括TTP(對等鏈路)標籤塊436、RX數據路徑432、RX存儲裝置432-1(例如，管芯上SRAM)、TX數據路徑434、和TX存儲裝置434-1(例如，管芯上SRAM)。硬件重放架構可以重放在諸如TTP之類的有損協議下傳輸期間丟失的分組。可選地，節點400的TTP媒體訪問控制(MAC)塊410可以進一步包括TTP MAC RDMA地址編碼塊438，其可以接收和編碼來自片上系統(SoC)420的RDMA發送數據418。

在一些實施例中，用於重放分組的節點400的硬件重放架構可以至少包括TTP標籤塊436、RX數據路徑432、RX存儲裝置432-1、TX存儲裝置434-1和TX數據路徑434的電路。如上面所討論的，硬件重放架構可以利用物理存儲裝置和數據結構來存儲在不同鏈路中傳輸和/或接收的分組，並維護傳輸的分組的次序，特別是當重放發生時。在一些實施例中，硬件重放架構所利用的物理存儲裝置可以是存儲、緩衝和/或保存與一個或多個鏈路相關聯的分組的任何合適類型的本地存儲裝置或高速緩存(例如，低級高速緩存)。物理存儲裝置的大小可能有限，諸如具有兆字節(MB)或千字節(KB)數量級的大小。在一些示例中，物理存儲裝置可以被部署為TX數據路徑434的一部

分，或者更具體地，作為TX存儲裝置434-1的一部分。物理存儲裝置也可以被部署為RX數據路徑432的一部分，或者更具體地，作為RX存儲裝置432-1的一部分。例如，物理存儲裝置可以是RX存儲裝置432-1和TX存儲裝置434-1，其中與RX數據路徑432和TX數據路徑434中的每一個相關聯的硬件重放架構所利用的RX存儲裝置432-1和TX存儲裝置434-1的大小可以是256 KB。在其他示例中，物理存儲裝置可以被部署在TTP標籤塊436內並作為TTP標籤塊436的一部分(例如，作為部署在TTP標籤塊436內的本地存儲裝置)。

應注意，節點400的TTP媒體訪問控制(MAC)塊410內的硬件重放架構可以採用任何其他合適大小的物理存儲裝置。在一些實施例中，硬件重放架構所利用的(例如，在TTP標籤塊436內的)數據結構可以包括一個或多個鏈表，其中每個鏈表可以記錄和/或跟蹤針對在第一通信節點和第二通信節點之間建立的對應鏈路所傳輸的分組的次序。在一些實施例中，TTP標籤塊436可以利用鏈表連同物理存儲裝置(例如，RX存儲裝置432-1和TX存儲裝置434-1)來維護和管理存儲的分組，以重放在多個鏈路上傳輸的分組。

圖8和圖9圖示了根據本公開的一些實施例的基於乙太網的網絡中的節點(例如，圖3B的節點400或設備A)所利用的示例物理存儲裝置和數據結構(例如，TX鏈表952)，所述基於乙太網的網絡實現了用於重放或重傳分組的TTP。

可以結合參考圖3B來理解圖8和圖9，圖3B示出了設備A響應於接收到通知沒有接收到分組(TTP_PAYLOAD ID=3)的未確認分組(例如，TTP_NACK ID=3)而重放兩個分組(例如，TTP_PAYLOAD ID=3到4)。

參考圖8，圖3B的設備A可以將用於傳輸和/或重放的分組1(例如，圖3B的分組TTP_PAYLOAD ID=1)、分組2(例如，圖3B的分組TTP_PAYLOAD ID=2)、分組3(例如，圖3B的分組TTP_PAYLOAD ID=3)、分組4(例如，圖3B的分組TTP_PAYLOAD ID=4)、分組5(例如，圖3B的分組TTP_CLOSE ID=5)存儲在物理存儲裝置(例如分組物理高速緩存802)中。如上所述，分組物理高速緩存802可以是TX存儲裝置434-1和/或可以是部署在TTP標籤塊436內的物理存儲裝置。在一些實施例中，分組物理高速緩存802可以具有兩個存儲空間——分組物理標籤804和分組物理數據806。對於分組(例如，分組1至分組5)中的每一個，分組物理標籤804可以包括物理地址指針，該物理地址指針指向存儲該分組的分組物理數據中的物理地址。例如，存儲在分組物理標籤804的條目中的與分組4相關聯的物理地址指針808可以指向存儲分組4(例如，圖3B的分組TTP_PAYLOAD ID=4)的分組物理數據806的條目。如圖8中所圖示，設備A可以按次序820傳輸分組1、分組2、分組3、分組4和分組5(例如，首先傳輸分組1並且最後傳輸分組5)。然而，設備A可以不基於次序820將分組1至分組5存儲在分組物理數據806中。具體地，儘管設備A在分組4和

分組 5 之前傳輸分組 3，但是存儲分組 3 的分組物理數據 806 中的地址 810 可以分別跟隨存儲分組 4 和分組 5 的分組物理數據 806 中的地址 812 和地址 814。

圖 9 圖示了 TX 鏈表 952，其可以由圖 3B 的節點 400 和 / 或設備 A 用來維護先前傳輸和重放之間的分組傳輸的次序。TX 鏈表 952 可以是節點 400 的 TTP 標籤塊 436 的一部分。如上面在討論圖 8 時所述，圖 3B 的設備 A 可以將分組 1 至分組 5 存儲在分組物理數據 806 的各個地址處，這些地址不反映分組 1 至分組 5 要被傳輸的次序 820。儘管如此，設備 A 也可以利用 TX 鏈表 952 來跟蹤並維護傳輸分組 1 至分組 5 的期望次序。如圖 9 中所示，TX 鏈表 952 包括五個元素 960、962、964、968、970，其中每個元素對應於或關聯於分組 1 至分組 5 之一。圖 9 圖示了 TX 鏈表 952 跟蹤並維護傳輸分組 1 至分組 5 的次序 820。例如，在 TX 鏈表 952 中，對應於分組 3 的元素 964 位於對應於分組 4 的元素 968 之前並指向元素 968，並且對應於分組 4 的元素 968 位於對應於分組 5 的元素 970 之前並指向元素 970。照此，通過利用 TX 鏈表 952，設備 A 可以在先前的傳輸和重放期間維護分組傳輸的次序，其中可以響應於接收到通知具有 TTP_PAYLOAD ID = 3 的分組沒有被圖 3B 中的設備 B 接收到的 TTP_NACK ID = 3 分組來觸發重放。根據本文中公開的任何合適的原理和優點，響應於超時或未確認，可以觸發重放。

如圖 9 中所示，圖 3B 的設備 A 可以進一步使用存儲在存儲器中的一個或多個指針 972、974 和 976 來確定要重放

的(一個或多個)分組。如圖3B的(3)和(4)中所圖示，設備A傳輸四個分組(例如，TTP_PAYLOAD ID=1至4)並接收三個分組(例如，TTP_ACK ID=1至2，以及TTP_NACK ID=3)，從而確認接收到由設備A傳輸的兩個分組(ID=1至2)，但是通知沒有接收到具有TTP_PAYLOAD ID=3的分組。作為響應，設備A可以將指針972設置為指向對應於分組3的元素964，以指示設備A要重放從分組3開始的分組。設備A可以進一步設置指針974指向對應於分組4的元素968，以指示設備A除了分組5之外還要重放分組4。設備可以進一步將指針976設置為指向對應於分組5的元素970，以指示設備A可以在重放分組3和分組4之後傳輸分組5。此外，設備A可以將TX鏈表952的元素960和元素962設置為空，以指示分組1和分組2可以從分組物理數據806和分組物理標籤804的地址(圖8中未示出)中移除，以釋放更多的存儲空間來存儲由設備A傳輸或接收的分組。

此後，基於TX鏈表952、指針972和指針974，設備A可以重放分組3和分組4，如圖3B的(5)中所圖示。然後，如圖3B的(6)中所圖示，設備A可以接收對接收到分組3和分組4的確認。作為響應，基於TX鏈表952，設備A可以傳輸與TX鏈表952的元素970相對應的分組5(例如，圖3B的分組TTP_CLOSEID=5)，以完成分組1至分組5的傳輸和重放。附加地和/或可選地，在對應於TX鏈表952的元素的的所有分組已經被傳輸和重放之後，設備A可以釋放分組1至分組5所佔用的存儲裝置。在一些實施例中，設備A可以通過

將空閒列表條目 832 和空閒列表條目 834 分別設置為特定值，來指示分組物理標籤 804 中的地址和分組物理數據 806 中的地址已經被釋放，並且可以結合對應於其他分組的 (一個或多個) 其他鏈表一起自由使用。

圖 10 圖示了根據本公開的一些實施例的圖 4 的 TTP 標籤塊 436 的示例框圖，其中，TTP 標籤塊 436 是用於重放在多個鏈路上傳輸的分組的硬件重放架構的一部分。如圖 10 中所示，TTP 標籤塊 436 可以包括存儲 TX 鏈表 1020 的存儲器和分別在流水線級 1002、1004、1006 和 1008 中操作的邏輯電路 1012、1014、1016 和 1018。邏輯電路 1012、1014、1016 和 1018 可以由任何合適的物理電路來實現。在一些示例中，邏輯電路 1012、1014、1016 和 1018 中的一些或全部可以由專用電路實現，諸如以專用集成電路 (ASIC) 的形式。在一些其他示例中，邏輯電路 1012、1014、1016 和 1018 中的一些或全部可以由可編程邏輯門或通用處理電路實現，諸如以現場可編程門陣列 (FPGA) 或數字信號處理器 (DSP) 的形式。在操作中，TX 鏈表 1020 可以類似於圖 9 的 TX 鏈表 952 運轉。在一些實施例中，TX 鏈表 1020 跟蹤包括分組 1022、分組 1024 和分組 1026 的 N 個分組的次序，其中節點 400 可以在特定鏈路上傳輸由 TX 鏈表 1020 跟蹤的 N 個分組。TTP 標籤塊 436 進一步包括分別指向分組 1022、分組 1024 和分組 1026 的指針 1032、指針 1034 和指針 1036。TTP 標籤塊 436 可以在任何合適的存儲元素 (圖 10 中未示出) 中存儲指針 1032、指針 1034 和指針 1036。在某些應用中，

包括TX鏈表1020的分組1022、分組1024和分組1026的N個分組可以存儲在物理存儲裝置中，諸如節點400的TX數據路徑434的TX存儲裝置434-1。在這樣的應用中，TX鏈表1020可以包括指向分組1022、1024、1026的指針。在其他應用中，包括分組1022、分組1024和分組1026的N個分組可以是存儲在TTP標籤塊436內的物理存儲裝置中的TX鏈表1020的一部分。

在一些實施例中，節點400可以將使用在TX存儲裝置434-1(或節點400的其它物理存儲裝置)中的TTP下建立的鏈路傳輸到第二節點的N個分組(包括分組1022、分組1024和分組1026)，N為可以受TX存儲裝置434-1的大小限制的任何正整數。只要來自TTP和/或網絡條件的約束准許，節點400可以不斷地向第二節點傳輸N個分組中的一些或全部。為了適應包括分組1022、分組1024和分組1026的N個分組的重放，TX存儲裝置434-1可以繼續存儲已經傳輸的一個或多個分組(例如，分組1022)，直到從第二節點接收到對接收到一個或多個分組的確認。可以存儲分組，直到確認接收到先前傳輸的分組為止。當接收到對接收到分組的確認時，TX存儲裝置434-1可以丟棄該分組，以騰出空間來存儲要在節點400和第二節點和/或一個或多個其他節點之間的鏈路或其他鏈路上傳輸的分組。相比之下，如果接收到分組的未確認(例如，第二節點通知節點400沒有接收到分組)或者在沒有從第二節點接收到對接收到分組的確認或未確認的情況下發生超時，則節點400可以重放仍

然存儲在TX存儲裝置434-1中的分組(例如，向第二節點重傳分組)。與重放分組相關聯，節點400可以丟棄已經接收到對接收的確認的其他分組。在一些實施例中，TX鏈表1020可以與TX存儲裝置434-1協調，以維護包括分組1022、分組1024和分組1026的N個分組中的一些或全部的先前傳輸與之後的任何重放之間的次序。如圖10中所示，TX鏈表1020包括N個元素，其中每個元素對應於或包括N個分組中的每一個以及對對應於下一個分組的下一個元素的引用。

當傳輸和/或重放N個分組時，TTP標籤塊436可以進一步利用分別指向TX鏈表1020中的三個元素的指針1032、指針1034和指針1036，以確定分組是要保留用於重放還是可以由TX存儲裝置434-1丟棄以節省存儲資源。取N為9(例如，從節點400傳輸到第二節點的9個分組)作為示例，在TX鏈表1020中，第1個元素對應於第1個分組(例如，分組1022)和第1個引用，其中第1個引用指向第2個元素；第2個元素對應於第2個分組和第2個引用，其中第2個引用指向第3個元素；並且第8個元素對應於第8個分組(例如，分組1024)和第8個引用，其中第8個引用指向第9個元素；並且第9個元素對應於第9個分組(例如，分組1026)。TTP標籤塊436可以維護和更新三個指針1032、1034和1036，這三個指針分別指向第1個元素(例如，分組1022)、第8個元素(例如，分組1024)和第9個元素(例如，分組1026)。

進一步假設節點400已經傳輸第1至第9個分組，並且已經從第二節點接收到對接收到第1至第7個分組但未接收到第8和第9個分組的確認，則指針1032然後指向TX鏈表1020的第1個元素(例如，分組1022)，指針1034然後指向TX鏈表1020的第8個元素(例如，分組1024)，並且指針1036然後指向TX鏈表1020的第9個元素(例如，分組1026)。照此，TTP標籤塊436可以使TX存儲裝置434-1丟棄包括分組1022、分組1024和分組1026的N個分組中的一些或全部，並且基於指針1032、1034和1036重放N個分組中的一些或全部。更具體地，TX存儲裝置434-1可以通過由指針1036指向的分組1026來重放由指針1034指向的分組1024(在這種情況下，僅重放分組1024和分組1026)。TX存儲裝置434-1可以進一步丟棄剩餘的分組(例如，由指針1032指向的分組1022以及在分組1024之前先前傳輸的其他分組；在這種情況下，可以丟棄包括分組1022的七個分組)。

如圖10中所圖示，TTP標籤塊436(例如，邏輯電路1012、1014、1016和1018)中的一些或全部可以以流水線方式操作，以增加節點400的吞吐量。邏輯電路1012、1014、1016和1018可以結合TX鏈表1020進行操作，以確定分組是否應該從TX存儲裝置434-1或存儲分組的節點400的其他物理存儲裝置中被重放或丟棄/退出。如圖10中所示，邏輯電路1012、1014、1016和1018可以根據TTP標籤塊436操作時的時鐘在相應的流水線級操作。具體地，邏

輯電路 1012 在初始流水線級 1002(標記為“Q0”)操作，邏輯電路 1014 在第一流水線級 1004(標記為“Q1”)操作，邏輯電路 1016 在第二流水線級 1006(標記為“Q2”)操作，並且邏輯電路 1018 在第三流水線級 1008(標記為“Q3”)操作。

在操作中，邏輯電路 1012 可以選擇數據流之一在 TTP 鏈路標籤流水線中進行處理。如初始流水線級 1002 中所示，邏輯電路 1012 可以基於控制信號(例如，“挑選”)選擇傳輸流(“TX 隊列”)、接收流(“RX 隊列”)或確認流(“ACK 隊列”)中的一個，用於在 TTP 鏈路標籤流水線中進行處理。在 TTP 鏈路標籤流水線中，邏輯電路確定是重放所選數據流的一個或多個分組，還是退出所選數據流的一個或多個分組。TTP 鏈路標籤流水線還可以確定拒絕對在 TTP 標籤流水線確定要重放的另一個分組之後傳輸的分組的確認。

假設邏輯電路 1012 選擇傳輸流以準備重放分組，則在第一流水線級 1004，邏輯電路 1014 確定評估哪個鏈路以進行重放。這可以涉及讀取與鏈路相關聯的標籤。如圖 10 中所示，邏輯電路 1014 可以選擇兩個鏈路(例如，“MOOSE”和“CAT”)中的一個用於可能的重放，其中每個鏈路可以建立在相同的端點或不同的端點之間。例如，可以在節點 400 和第二節點之間建立鏈路“MOOSE”和“CAT”兩者；替代地，可以在節點 400 和第二節點之間建立鏈路“MOOSE”，而在節點 400 和第三節點之間建立鏈路“CAT”。邏輯電路 1014 可以基於指向所選鏈路的鏈路指針來選擇用於重放的鏈路(例如，“CAT”)。

然後，在第二流水線級1006處，邏輯電路1016可以確定在鏈路“CAT”上傳輸的哪個(哪些)分組將被重放或退出。在一些實施例中，邏輯電路1016基於是否已經接收到對接收的確認或未確認，確定重放在鏈路“CAT”上傳輸的一些分組，而其他分組可以退出。例如，如果接收到對分組1024的未確認的接收或者在觸發超時的時間段內沒有接收到對分組1024的確認，則邏輯電路1016可以確定重放分組1024。相比之下，邏輯電路1016可以響應於接收到對分組1022的確認，確定退出分組1022。附加地和/或可選地，邏輯電路1016可以進一步基於TX鏈表1020來確定重放和/或退出在鏈路“CAT”上傳輸的其他分組。例如，基於由TX鏈表1020指定的鏈路“CAT”上傳輸的分組的次序(其示出分組1026是在分組1024之後傳輸的)，響應於接收到對分組1024的未確認，邏輯電路1016可以確定重放分組1026連同重放分組1024。假設已經接收到在分組1022和分組1024之間傳輸的對分組的確認，邏輯電路1016可以進一步使TX存儲裝置434-1退出在分組1022和分組1024之間傳輸的分組，以在TX存儲裝置434-1中騰出更多可用的存儲空間。在第二流水線級1006中，對分組的確認可以與確定重放較早傳輸的分組相關聯地被拒絕。退出分組可以涉及允許其他數據代替該分組被寫入存儲器和/或從存儲器中刪除該分組。

此後，在第三流水線級1008，邏輯電路1018可以更新指向鏈路“CAT”的鏈路指針，以指向另一鏈路(例如，鏈路

“MOOSE”)。照此，在下一輪流水線操作中，邏輯電路 1012、1014、1016和 1018可以基於另一 TX鏈表(圖 10中未示出)來確定是否重放與鏈路“MOOSE”相關聯的(一個或多個)分組，該另一 TX鏈表包括、引用或對應於在鏈路“MOOSE”上傳輸的分組。有利地，使用 TX存儲裝置 434-1 和 TX鏈表 1020來實現重放功能使得節點 400能夠在有限的硬件資源下使用 TTP與第二節點通信，而無需軟件控制機制的輔助。

硬件鏈路定時器

圖 11圖示了硬件鏈路定時器 1100的示例框圖，其實現超時檢查機制，用於在沒有軟件輔助的情況下重放分組。在一些實施例中，硬件鏈路定時器 1100可以是圖 4的節點 400的一部分。硬件鏈路定時器 1100中的一些或全部可以部署在圖 4的 TTP標籤塊 436內。如上所述，與其他乙太網協議(例如，TCP或UDP)形成對照——軟件通常採用該其他乙太網協議使用多個定時器(例如，一個定時器用於一個鏈路)來跟蹤多個鏈路上的超時——硬件鏈路定時器 1100可以允許節點 400確定在哪個(哪些)鏈路上傳輸的哪個(哪些)分組要重放，並且如果期望重放，則在有限的硬件資源下何時重放(例如，當虛擬和/或物理地址空間和計算資源的大型資源池不可用時)。在一些實施例中，硬件鏈路定時器 1100可以週期性地對由節點 400用來依照 TTP與一個或多個其他節點通信的已建立鏈路(例如，活動鏈路)執行時

序檢查。

如圖 11 中所示，硬件鏈路定時器 1100 可以包括先進先出 (FIFO) 存儲器 1104、定時器 1102 和邏輯電路 1120、1112、1114、1116 和 1118，其中邏輯電路 1112、1114、1116 和 1118 可以為用於重放分組的 TTP 標籤塊 436 的一部分。FIFO 存儲器 1104 可以存儲與每個活動鏈路相關聯的時序和狀態信息。硬件鏈路定時器 1100 可以以輪詢方式檢查存儲在 FIFO 存儲器 1104 中的與每個活動鏈路相關聯的時序和狀態。更具體地，硬件鏈路定時器 1100 可以開始針對存儲在 FIFO 存儲器 1104 的第 N 個條目中的與第 N 個鏈路相關聯的時序和狀態信息，檢查存儲在 FIFO 存儲器 1104 的第一個條目中的與第一個鏈路相關聯的時序和狀態信息，並且然後再次檢查存儲在 FIFO 存儲器 1104 的第一個條目中的與第一個鏈路相關聯的時序和狀態信息。硬件鏈路定時器 1100 可以利用定時器 1102 來調度時間點，以讀出與多個活動鏈路和 / 或分組相關聯的時序和狀態信息。讀出的時序和狀態信息可以用於通過進一步的信息查找來確定是重放與鏈路相關聯的分組還是退出和 / 或丟棄分組。應當注意，圖 4 的節點 400 可以包括類似於圖 11 中所圖示的多於一個硬件鏈路定時器，其中每個硬件鏈路定時器可能能夠確定是否存在與多個鏈路相關聯的超時。

在一些實施例中，FIFO 存儲器 1104 可以存儲與節點 400 和 (一個或多個) 其他節點之間建立的一個或多個鏈路相關聯的時序信息。例如，節點 400 可以包括硬件鏈路定

時器 1100，其使用 FIFO 存儲器 1104 來存儲與在節點 400 和一個或多個其他節點之間建立的 M 個鏈路相關聯的時序信息，其中 M 是大於一的正整數。代替使用 M 個定時器（其中每個定時器跟蹤對應鏈路的時序信息），硬件鏈路定時器 1100 可以利用定時器 1102（例如，在可編程時間段內滴答一次的硬件時鐘）通過以輪詢（例如，循環）方式訪問 FIFO 存儲器 1104 來跟蹤和/或更新 M 個鏈路中的每一個的時序信息。具體地，當定時器 1102 滴答一次時，硬件鏈路定時器 1100 可以以輪詢方式一次一個地訪問 FIFO 存儲器 1104 的條目，其中 FIFO 存儲器 1104 的每個被訪問的條目對應於 M 個鏈路之一。

在一些實施例中，定時器 1102 的每次滴答的時間段可以變化，並且可以在數百微秒至一位數微秒之間的數量級。例如，定時器 1102 的滴答的時間段可以高達 100 微秒，並且可以低至 1 微秒。此外，硬件鏈路定時器 1100 可以基於由 FIFO 存儲器 1104 的條目表示的鏈路數量（例如， M ）來調整定時器 1102 的滴答的時間段。例如，當 M 增加（例如，FIFO 存儲器 1104 的條目表示更多的鏈路）時，定時器 1102 的滴答的時間段可以減少；並且當 M 減小（例如，FIFO 存儲器 1104 的條目表示更少的鏈路）時，定時器 1102 的滴答的時間段可以增加。照此，如果定時器 1102 的滴答的時間段與由 FIFO 存儲器 1104 的條目所表示的鏈路數量不成比例地改變，則檢查鏈路的狀態和/或時序信息的時間間隔可以保持不變。

在一些實施例中，與M個鏈路之一相關聯的時序和/或狀態信息可以指示該鏈路已經有多長時間未接收到被傳輸的對接收到分組的確認。假設節點400已經在鏈路上向第二節點傳輸了N個分組，FIFO存儲器1104的一個條目可以存儲時序和/或狀態信息，所述時序和/或狀態信息當在定時器1102的滴答的特定時間段下通過輪詢方式被訪問時指示在預確定的持續時間(例如，20微秒、50微秒、100微秒、200微秒、300微秒、400微秒、500微秒、和/或其間的任何持續時間)內沒有接收到對接收到N個分組中的任何一個的確認。在訪問FIFO存儲器1104的條目時，硬件鏈路定時器1100可以利用邏輯電路1120、1112、1114、1116和1118來檢查存儲在條目中的時序和/或狀態信息，並且查找可以存儲在節點400的本地存儲裝置(例如，TX存儲裝置434-1或其他本地存儲裝置)中的N個分組，以用於重放N個分組。

替代地，可以將與M個鏈路之一相關聯的時序和/或狀態信息存儲在FIFO存儲器1104的一個條目中，以指示該鏈路可以關閉(例如，由第一節點傳輸的所有分組已經被第二節點接收到)。在訪問FIFO存儲器1104的條目時，硬件鏈路定時器1100可以利用邏輯電路1120、1112、1114、1116和1118來檢查存儲在條目中的時序和/或狀態信息，並且查找可能仍然存儲在節點400的本地存儲裝置(例如，TX存儲裝置434-1)中的分組，並且丟棄該分組，因為存儲在FIFO存儲器1104的條目中的時序和/或狀態信息指示該

鏈路可以被關閉。有利地，通過利用在多個鏈路和/或分組的可調整週期下滴答的單個定時器(例如，定時器1102)和存儲多個鏈路的時序和/或狀態信息的FIFO存儲器1104，節點400可以以適當的時序重放分組，以實現低時延並釋放由非活動鏈路(例如，關閉的鏈路)佔用的硬件資源，以供活動鏈路使用，從而在有限的計算和存儲資源下操作。

如圖11中所圖示，邏輯電路1120、1112、1114、1116和1118可以在不同的流水線級中操作，類似於圖10中所圖示的邏輯電路1012、1014、1016和1018。如圖11中所示，邏輯電路1120、1112、1114、1116和1118可以結合定時器1102和FIFO存儲器1104進行操作，以確定何時需要重放在一個或多個鏈路上傳輸的分組，或者何時可以從本地存儲裝置(諸如TX存儲裝置434-1)中退出/丟棄該分組，或者是否可以關閉所述一個或多個鏈路。如圖11中所示，邏輯電路1120、1112、1114、1116和1118可以根據硬件鏈路定時器1100操作時的時鐘在相應的流水線級操作。具體地，邏輯電路1120和1112可以在初始流水線級(標記為“Q0”)操作，邏輯電路1114可以在第一流水線級(標記為“Q1”)操作，邏輯電路1116可以在第二流水線級(標記為“Q2”)操作，邏輯電路1118可以在第三流水線級(標記為“Q3”)操作。

在操作中，在初始流水線級Q0，邏輯電路1120可以選擇用於邏輯電路1112的時序和狀態信息查找(例如，定時

器鏈路查找)的時序和狀態信息。如圖 11 中所示，時序和狀態信息可以來自來自 FIFO 存儲器 1104 的條目(例如，比所有其他條目更早進入 FIFO 存儲器 1104 的最老的條目)，或者來自其他來源(例如，替代的優先級鏈路查找信息)。如圖 11 中所圖示，在初始流水線級 Q0，邏輯電路 1112 基於選擇“定時器鏈路查找”而不是“TX 業務”或“RX 業務”的控制信號(例如，“挑選”)來選擇與 FIFO 存儲器 1104 中的“鏈路 A”相關聯的時序和狀態信息。“TX 業務”可以對應於在由節點 400 建立的鏈路(例如，“鏈路 B”)上傳輸的分組，而“RX 業務”可以對應於在由節點 400 建立的另一鏈路(例如，“鏈路 D”)上接收的分組。

在第一流水線級 Q1，邏輯電路 1114 基於從初始流水線級 Q0 接收的時序和狀態信息，確定正在查詢哪個鏈路。如圖 11 中所圖示，邏輯電路 1114 確定“鏈路 A”正在被查詢，以便稍後確定“鏈路 A”是否需要被重放或者是否可以被關閉。然後，在第二流水線級 Q2，邏輯電路 1116 基於從 FIFO 存儲器 1104 訪問的與“鏈路 A”相關聯的時序和狀態信息，確定“鏈路 A”是否可以被關閉。如果與“鏈路 A”相關聯的時序和狀態信息示出“鏈路 A”可以關閉，則邏輯電路 1116 可以觸發與“鏈路 A”相關聯的分組從本地存儲裝置(例如，TX 存儲裝置 434-1)中退出/丟棄。如果與“鏈路 A”相關聯的時序和狀態信息示出“鏈路 A”仍然是活動的/打開的，則硬件鏈路定時器 1100 的操作進行到第三流水線級 Q3，其中邏輯電路 1118 確定是否重放在“鏈路 A”上傳輸的分組，

或者如何更新與“鏈路 A”相關聯的時序和狀態信息。

在第三流水線級 Q3，邏輯電路 1118 可以基於從 FIFO 存儲器 1104 中訪問的與“鏈路 A”相關聯的狀態和時序信息，確定重放與“鏈路 A”相關聯的至少一些分組。例如，與“鏈路 A”相關聯的狀態和時序信息可以包括“定時器比特”，該“定時器比特”在被設置(例如，設置為邏輯 1)時可以指示節點 400 在用於重放分組的閾值持續時間內沒有接收到對接收到與“鏈路 A”相關聯的分組中的至少一個分組的確認。在一些實施例中，閾值持續時間可以是可調整的，並且可以是 20 微秒、50 微秒、100 微秒、200 微秒、300 微秒、400 微秒、500 微秒、和/或其間的任何合適的持續時間。閾值持續時間可以在從 20 微秒到 500 微秒的範圍內。在一些實施例中，可以基於已經從 FIFO 存儲器 1104 查詢“鏈路 A”的次數和定時器 1102 的時間段來設置與“鏈路 A”(和/或其他鏈路)相關聯的“定時器比特”。

如果“定時器比特”被斷言，則邏輯電路 1118 可以使與“鏈路 A”相關聯的分組被重放。被斷言的“定時器比特”可以指示與一個或多個分組相關聯的超時已經發生(例如，在沒有接收到確認或未確認的情況下已經達到閾值持續時間)。此外，邏輯電路 1118 可以響應於“鏈路 A”的重放，更新存儲在 FIFO 存儲器 1104 中的與“鏈路 A”相關聯的時序和狀態信息。例如，邏輯電路 1118 可以清除“定時器比特”(例如，將“定時器比特”從邏輯 1 設置為邏輯 0)。另一方面，如果與“鏈路 A”相關聯的狀態和時序信息指示不重放

“鏈路 A”上的一個或多個分組(例如，“定時器比特”沒有被斷言，這對應於是圖 11 中的邏輯 0)，則邏輯電路 1118 可以不使“鏈路 A”被重放。在這樣的情況下，如果與“鏈路 A”相關聯的時序和狀態信息指示“鏈路 A”在下一次被查詢的情況下應該被重放，則邏輯電路 1118 可以進一步將“定時器比特”設置為邏輯 1。

重放和鏈路時序的示例方法

現在轉向圖 12，將描述用於重放從節點(諸如圖 3B 的節點 400 或設備 A)傳輸的分組的說明性分組重放過程 1200。分組重放過程 1200 可以例如由圖 4 的節點 400 的 TTP 標籤塊 436 或其他組件來實現。過程 1200 開始於框 1202，其中 TTP 標籤塊 436 可以存儲包括使用乙太網協議在第一鏈路上從節點 400 傳輸到第二節點的分組的鏈表。例如，鏈表可以是 TX 鏈表 1020，其包括或引用分組 1022、1024 和 1026，以維護分組 1022、1024 和 1026 向第二節點傳輸的次序。

在框 1204 處，TTP 標籤塊 436 可以確定響應於以下中的至少一項來重放所述分組中的第一分組：(a) 從第二節點接收到對第一分組的未確認或 (b) 與第一分組相關聯的超時。例如，TTP 標籤塊 436 可以響應於 (a) 從第二節點接收到對分組 1024 的未確認或者 (b) 與分組 1024 相關聯的超時，指示在閾值時間段內沒有接收到對分組 1024 的確認，來確定重放分組 1024。

在框 1206 處，TTP 標籤塊 436 可以響應於從第二節點接收到對第二分組的確認，退出所述分組中的第二分組。例如，響應於從第二節點接收到對分組 1022 的確認，TTP 標籤塊 436 可以退出分組 1022。

圖 13 圖示了用於確定是否重放與節點 (諸如圖 3B 的節點 400 或設備 A) 相關聯的一個或多個鏈路的示例鏈路超時過程 1300。鏈路超時過程 1300 可以例如由圖 11 的硬件鏈路定時器 1100 或節點 400 來實現。過程 1300 開始於框 1302，其中硬件鏈路定時器 1100 或節點 400 在 FIFO 存儲器中存儲與多個鏈路相關聯的時序和狀態信息，並且節點 400 使用乙太網協議在多個鏈路上向一個或多個其他節點傳輸分組。例如，硬件鏈路定時器 1100 可以在 FIFO 存儲器 1104 中存儲與多個鏈路相關聯的時序和狀態信息。

在框 1304 處，硬件鏈路定時器 1100 或節點 400 可以基於硬件鏈路定時器 1100 或節點 400 內部署的硬件定時器的相應滴答來訪問 FIFO 存儲器的條目。例如，硬件鏈路定時器 1100 可以基於定時器 1102 的相應滴答來訪問 FIFO 存儲器 1104 的條目。

在框 1306 處，硬件鏈路定時器 1100 或節點 400 可以基於與多個鏈路中的第一鏈路相關聯的時序和狀態信息，確定重放與第一鏈路相關聯的至少一個分組。例如，硬件鏈路定時器 1100 可以基於與“鏈路 A”相關聯的時序和狀態信息來確定重放與“鏈路 A”相關聯的或者在“鏈路 A”上傳輸的至少一個分組。

結 論

前述公開內容無意將本公開內容限制在所公開的精確形式或特定使用領域。照此，設想的是，鑒於本公開，對本公開的各種替代實施例和/或修改——無論是在本文中明確描述的還是暗示的——都是可能的。已經如此描述了本公開的實施例，本領域普通技術人員將認識到，在不脫離本公開的範圍的情況下，可以在形式和細節上進行改變。因此，本公開僅受申請專利範圍的限制。

應理解，根據本文中描述的任何特定示例，不一定可以實現所有目的或優點。因此，例如，本領域技術人員將認識到，一些示例可以以實現或優化如本文中教導的一個優點或一組優點的方式操作，而不一定實現可以如本文中教導或建議的其他目的或優點。

本文中描述的所有過程可以經由由包括計算機或處理器的計算系統執行的軟件代碼模塊來體現，並完全自動化。代碼模塊可以存儲在任何類型的非暫時性計算機可讀介質或其他計算機存儲設備中。一些或所有方法可以在專用計算機硬件中體現。

除本文中描述之外的許多其他變型將從本公開中是顯而易見的。例如，取決於該示例，本文中描述的任何算法的一些動作、事件或功能可以以不同的順序執行，可以被添加、合併或一起省略(例如，並非所有描述的動作或事件對於算法的實踐都是必要的)。此外，在一些示例中，

動作或事件可以例如通過多線程處理、中斷處理、或多個處理器或處理器核同時執行或者在其他並行架構上同時執行，而不是順序執行。此外，不同的任務或過程可以由可以一起運轉的不同機器和/或計算系統來執行。

結合本文中公開的示例所描述的各種說明性邏輯塊和模塊可以由機器實現或執行，該機器諸如處理單元或處理器、數字信號處理器(DSP)、專用集成電路(ASIC)、現場可編程門陣列(FPGA)或其他可編程邏輯設備、離散門或晶體管邏輯、離散硬件組件、或設計用於執行本文中描述功能的任何組合。處理器可以是微處理器，但是替代地，處理器可以是控制器、微控制器或狀態機、它們的組合等。處理器可以包括處理計算機可執行指令的電路。在一些示例中，處理器包括執行邏輯操作而不處理計算機可執行指令的FPGA或其他可編程設備。處理器也可以被實現為計算設備的組合，例如，DSP和一個微處理器、多個微處理器、結合DSP核的微處理器、或任何其他這樣的配置的組合。儘管本文中主要針對數字技術進行了描述，但是處理器也可以主要包括模擬組件。計算環境可以包括任何類型的計算機系統，包括但不限於基於微處理器的計算機系統、大型計算機、數字信號處理器、便攜式計算設備、設備控制器、或器具內的計算引擎等等。

結合本文中公開的實施例描述的方法、過程、例程或算法的元素可以直接體現在硬件中、由處理器設備執行的軟件模塊中、或兩者的組合中。軟件模塊可以駐留在RAM

存儲器、閃速存儲器、ROM存儲器、EPROM存儲器、EEPROM存儲器、寄存器、硬盤、可移除盤、CD-ROM、或任何其他形式的非暫時性計算機可讀存儲介質中。示例性存儲介質可以耦合到處理器設備，使得處理器設備可以從存儲介質讀取信息以及向存儲介質寫入信息。作為替代，存儲介質可以集成到處理器設備中。處理器設備和存儲介質可以駐留在ASIC中。ASIC可以駐留在用戶終端中。作為替代，處理器設備和存儲介質可以作為分立組件駐留在用戶終端中。

本文中描述或本公開圖式中所圖示的過程可以響應於事件而開始，諸如根據預確定或動態確定的時間表而開始、根據由用戶或系統管理員啟動時的要求而開始、或響應於某個其他事件而開始。當啟動這樣的過程時，存儲在一個或多個非暫時性計算機可讀介質(例如，硬盤驅動器、閃速存儲器、可移除介質等)上的一組可執行程序指令可以被加載到服務器或其他計算設備的存儲器(例如，RAM)中。然後，可執行指令可以由計算設備的基於硬件的計算機處理器來執行。在一些實施例中，這樣的過程或其部分可以串行或並行地在多個計算設備和/或多個處理器上實現。

除其他之外，諸如“可以”(“can”、“could”)或“可能”(“might”、“may”)的條件語言，除非另有具體陳述，否則在上下文內理解為一般用於傳達一些示例包括一些特徵、元素和/或步驟，而其他示例不包括一些特徵、元素

和/或步驟。因此，這樣的條件語言一般不意圖暗示特徵、元素和/或步驟以任何方式作為示例，或者示例一定包括用於在有或沒有用戶輸入或提示的情況下決定這些特徵、元素和/或步驟是否被包括在任何特定示例中或者是否要在任何特定示例中執行的邏輯。

除非另有具體陳述，否則諸如短語“X、Y或Z中的至少一個”的析取語言被理解為如一般用於呈現項目、術語等可以是X、Y或Z或其任何組合(例如，X、Y和/或Z)的上下文。因此，這樣的析取語言一般不意圖並且不應該暗示一些示例需要X中的至少一個、Y中的至少一個、或Z中的至少一個各自存在。

本文中描述和/或圖式中所描繪的流程圖中的任何過程描述、元素或框應理解為潛在表示模塊、代碼段或代碼部分，其包括用於實現過程中具體邏輯功能或元素的可執行指令。替代示例被包括在本文中描述的示例的範圍內，其中元素或功能可以被刪除、不按所示或所討論的次序(包括基本上同時或按相反次序，這取決於所涉及的功能)執行，如本領域技術人員將理解的。

應強調的是，可以對上述示例進行許多變型和修改，其元素應被理解為除其他之外可接受的示例。所有這樣的修改和變型都意圖被包括在本公開的範圍內。

本文中描述和/或圖式中所描繪的流程圖中的任何過程描述、元素或框應理解為潛在表示模塊、代碼段或代碼部分，其包括用於實現過程中具體邏輯功能或元素的可執

行指令。替代實施方式被包括在本文中描述的示例的範圍內，其中元素或功能可以被刪除、不按所示或所討論的次序(包括基本上同時或按相反次序，這取決於所涉及的功能)執行，如本領域技術人員將理解的。

除非另有明確陳述，否則諸如“一”或“一個”的冠詞一般應解釋為包括一個或多個所描述的項目。因此，諸如“設備被配置為”的短語意圖包括一個或多個所述設備。這樣一個或多個所述設備也可以被共同配置來實行所述敘述。例如，“處理器被配置為實行敘述 A、B 和 C”可以包括被配置為實行敘述 A 的第一處理器結合被配置為實行敘述 B 和 C 的第二處理器一起工作。

【符號說明】

202:關閉

204:打開接收

206:打開發送

208:打開

210:關閉接收

212:關閉發送

400:節點

402:PCS+PMA IEEE802.3

404:156.25 MHz的參考時鐘

408:RX幀

410:TTP MAC塊

- 412:TX 幀
- 416:RDMA 接收 數據
- 418:RDMA 發 送 數據
- 420:SoC NOC+存 儲 器
- 422:TTP FSM
- 432-1:RX 管 芯 上 SRAM 暫 時 性 存 儲 裝 置
- 432:RX 數 據 路 徑
- 434:TX 數 據 路 徑,重 放 數 據 路 徑
- 434-1:TX 管 芯 上 SRAM 暫 時 性 存 儲 裝 置
- 436:TTP 對 等 鏈 路 標 簽
- 438:TTP MAC(可 選 IP)RDMA 地 址 編 碼
- 600:網 絡 和 計 算 環 境
- 602A~602E:主 機
- 604A~604N:主 機
- 606A~606N:圖 塊
- 608:乙 太 網 交 換 機
- 610:端 口
- 802:分 組 物 理 高 速 緩 存
- 804:分 組 物 理 標 簽
- 806:分 組 物 理 數 據
- 808:物 理 地 址 指 針
- 810:地 址
- 812:地 址
- 814:地 址

- 820:次序
- 832:空閒列表
- 834:空閒列表
- 952:TX鏈表
- 960:ID=1 PA=null(ACK)
- 962:ID=2 PA=null(ACK)
- 964:ID=3 PA=15
- 968:ID=4 PA=2
- 970:ID=5 PA=3
- 972:退出指針
- 974:當前指針
- 976:分配指針
- 1002:Q0
- 1004:Q1
- 1006:Q2
- 1008:Q3
- 1012:邏輯電路
- 1014:邏輯電路
- 1016:邏輯電路
- 1018:邏輯電路
- 1020:TX鏈表
- 1022:OLD CAT
- 1024:NEWER CAT
- 1026:CAT

- 1032:退出指針
- 1034:當前指針
- 1036:分配指針
- 1100:硬件鏈路定時器
- 1102:定時器
- 1104:FIFO存儲器
- 1112:邏輯電路
- 1114:邏輯電路
- 1116:邏輯電路
- 1118:邏輯電路
- 1120:邏輯電路
- 1200:步驟
- 1202:步驟
- 1204:步驟
- 1206:步驟
- 1208:步驟
- 1300:步驟
- 1302:步驟
- 1304:步驟
- 1306:步驟
- 1308:步驟

【發明申請專利範圍】

【請求項1】一種用於基於乙太網的通信的第一節點，所述第一節點包括：

一個或多個處理器，其被配置為實現傳輸層僅硬件乙太網協議。

【請求項2】根據請求項1所述的第一節點，其中所述乙太網協議是有損的。

【請求項3】根據請求項2所述的第一節點，其中所述一個或多個處理器進一步被配置為實現硬件重放架構，以重放在第一鏈路上傳輸到第二節點的分組，其中所述分組存儲在第一節點的本地存儲裝置中，並且其中用於重放的分組的次序在鏈表中指定。

【請求項4】根據請求項1所述的第一節點，其中所述第一節點被配置為以一位數微秒時延向第二節點傳輸分組。

【請求項5】根據請求項1所述的第一節點，其中所述一個或多個處理器被配置為實現狀態機，所述狀態機被配置為：

在第一節點和第二節點之間的鏈路打開的打開狀態下操作；

從打開狀態轉換到中間關閉狀態；和

響應於從第二節點接收到關閉確認，從中間關閉狀態轉換到關閉狀態以關閉鏈路。

【請求項6】根據請求項1所述的第一節點，進一步包

括乙太網端口。

【請求項7】根據請求項1所述的第一節點，其中所述一個或多個處理器被配置為基於存儲在先進先出(FIFO)存儲器中的與鏈路相關聯的時序和狀態信息來確定在第一節點和第二節點之間的鏈路上重放分組，其中根據與多個鏈路相關聯的硬件鏈路定時器的滴答來訪問FIFO存儲器的條目。

【請求項8】一種用於基於乙太網的通信的第一節點，所述第一節點包括：

一個或多個處理器，其被配置為實現第2層僅硬件乙太網協議。

【請求項9】根據請求項8所述的第一節點，其中所述一個或多個處理器包括僅硬件架構，所述僅硬件架構被配置為對在第一鏈路上傳輸到第二節點的分組進行重放。

【請求項10】根據請求項8所述的第一節點，其中所述一個或多個處理器進一步被配置為基於存儲在先進先出(FIFO)存儲器中的與關聯於第一節點的鏈路相關聯的時序和狀態信息來確定在與所述第一節點相關聯的鏈路上重放分組，所述先進先出(FIFO)存儲器是基於與多個鏈路相關聯的定時器的滴答來訪問的。

【請求項11】一種第一節點，其中所述第一節點被配置為打開和關閉與基於乙太網的網絡中的第二節點的鏈路，所述第一節點包括：

狀態機硬件，其被配置為：

在第一節點和第二節點之間的鏈路打開的打開狀態下操作；

從打開狀態轉換到中間關閉狀態；和

響應於從第二節點接收到關閉確認，從中間關閉狀態轉換到關閉狀態以關閉鏈路，

其中所述第一節點被配置為在有損網絡中操作。

【請求項12】根據請求項11所述的第一節點，其中所述狀態機硬件以僅硬件實現傳輸層的流控制協議。

【請求項13】根據請求項12所述的第一節點，其中與流控制協議相關聯的時延小於10微秒。

【請求項14】根據請求項11所述的第一節點，其中所述狀態機硬件被配置為：

從關閉狀態轉換到中間打開狀態；和

從中間打開狀態轉換到打開狀態。

【請求項15】根據請求項11所述的第一節點，其中所述狀態機硬件響應於向第二節點傳輸關閉鏈路的請求或者從第二節點接收到關閉鏈路的請求，從打開狀態轉換到中間關閉狀態。

【請求項16】根據請求項11所述的第一節點，其中所述狀態機硬件響應於向第二節點傳輸關閉鏈路的確認，從中間關閉狀態轉換到關閉狀態。

【請求項17】根據請求項11所述的第一節點，其中所述狀態機硬件從中間關閉狀態轉換到關閉狀態，而不等待一段時間。

【請求項18】根據請求項11所述的第一節點，其中在打開狀態下，第一節點不重傳分組，直到從第二節點接收到對分組的未確認，或者在沒有接收到對分組的未確認的情況下預確定的超時時段到期為止。

【請求項19】根據請求項11所述的第一節點，其中在打開狀態下，第一節點不暫停地傳輸最多N個分組，並且其中N受到分配給第一節點的物理存儲器的大小限制。

【請求項20】根據請求項11所述的第一節點，進一步包括：

與多個鏈路相關聯的硬件鏈路定時器；和

硬件重放架構，其被配置為以僅硬件重放分組。

【請求項21】一種第一節點，包括：

硬件重放架構，其被配置為重放使用乙太網協議在第一鏈路上傳輸到第二節點的分組，其中所述硬件重放架構包括：

本地存儲裝置，其被配置為存儲包括分組的鏈表，其中所述鏈表維護用於向第二節點傳輸的分組的次序；和

邏輯電路，其被配置為：

響應於(a)從第二節點接收到對第一分組的未確認或者(b)與第一分組相關聯的超時中的至少一個，確定重放分組中的第一分組，和

響應於從第二節點接收到對第二分組的確認，退出分組中的第二分組，

其中所述乙太網協議是有損的。

【請求項22】根據請求項21所述的第一節點，其中所述邏輯電路包括多個流水線級，並且其中所述邏輯電路確定在所述多個流水線級中的第一流水線級處處理與第一節點和第二節點之間的第一鏈路而非第二鏈路相關聯的數據。

【請求項23】根據請求項22所述的第一節點，其中所述邏輯電路確定在所述多個流水線級中的第二流水線級重放第一分組。

【請求項24】根據請求項23所述的第一節點，其中所述邏輯電路在所述多個流水線級中的第二流水線級處，基於由鏈表維護的分組的次序，確定重放分組中的第三分組和分組中的第一分組。

【請求項25】根據請求項22所述的第一節點，其中所述邏輯電路基於鏈路指針確定處理與第一鏈路而非第二鏈路相關聯的數據，並且其中所述邏輯電路更新鏈路指針以指向所述多個流水線級中的第三流水線級處的第二鏈路。

【請求項26】根據請求項21所述的第一節點，其中所述第一節點和第二節點處於基於乙太網的網絡中，並且其中所述第一節點通過乙太網交換機與第二節點通信。

【請求項27】根據請求項26所述的第一節點，其中所述第一節點包括網絡接口處理器(NIP)和高帶寬存儲器(HBM)，並且其中所述HBM的帶寬至少為一千兆字節。

【請求項28】一種用於基於乙太網的通信的第一節

點，所述第一節點包括：

一個或多個處理器，其被配置為實現傳輸層僅硬件乙太網協議，

其中所述傳輸層僅硬件乙太網協議是有損的，並且

其中所述一個或多個處理器包括硬件重放架構，所述硬件重放架構被配置為重放在傳輸層僅硬件乙太網協議下傳輸的分組。

【請求項29】根據請求項28所述的第一節點，其中所述硬件重放架構包括：

本地存儲裝置，其被配置為存儲在傳輸層僅硬件乙太網協議下傳輸的分組。

【請求項30】根據請求項29所述的第一節點，其中所述硬件重放架構包括：

鏈表，其存儲在本地存儲裝置中，並且被配置為跟蹤用於傳輸到另一節點的分組的次序，其中鏈表的每個元素對應於存儲在本地存儲裝置中的每個分組。

【請求項31】根據請求項30所述的第一節點，其中所述硬件重放架構被配置為以對應於鏈表的次序傳輸分組。

【請求項32】根據請求項30所述的第一節點，其中所述硬件重放架構被配置為存儲：

第一指針，其被配置為指向鏈表的第一元素，其中所述第一指針指示不重放與鏈表的第一元素相對應的分組中的第一分組；和

第二指針，其被配置為指向鏈表的第二元素，其中所

述第二指針指示重放與鏈表的第二元素相對應的分組中的第二分組。

【請求項33】根據請求項32所述的第一節點，其中所述硬件重放架構根據用於傳輸的分組的次序重放第二分組以及第二分組之後的一個或多個分組。

【請求項34】根據請求項33所述的第一節點，其中所述硬件重放架構使本地存儲裝置根據用於傳輸的分組的次序丟棄第一分組和在第二分組之前的一個或多個分組。

【請求項35】一種在第一節點處實現的計算機實現的方法，用於重放使用乙太網協議在第一鏈路上傳輸到第二節點的分組，所述計算機實現的方法包括：

存儲包括分組的鏈表，其中所述鏈表維護用於向第二節點傳輸的分組的次序；

響應於(a)從第二節點接收到對第一分組的未確認或者(b)與第一分組相關聯的超時中的至少一個，確定重放分組中的第一分組；和

響應於從第二節點接收到對第二分組的確認，退出分組中的第二分組，

其中所述乙太網協議是有損的。

【請求項36】根據請求項35所述的計算機實現的方法，其中所述第一節點包括硬件重放架構，所述硬件重放架構包括多個流水線級，並且其中所述硬件重放架構確定在所述多個流水線級中的第一流水線級處處理與第一鏈路而非第二鏈路相關聯的數據。

【請求項37】根據請求項36所述的計算機實現的方法，其中所述硬件重放架構確定在所述多個流水線級中的第二流水線級重放第一分組。

【請求項38】根據請求項37所述的計算機實現的方法，其中所述硬件重放架構基於由所述多個流水線級中的第二流水線級處的鏈表維護的分組的次序，確定重放分組中的第三分組和分組中的第一分組。

【請求項39】根據請求項35所述的計算機實現的方法，其中所述第一節點和第二節點處於基於乙太網的網絡中，並且其中所述第一節點通過乙太網交換機與第二節點通信。

【請求項40】根據請求項39所述的計算機實現的方法，其中所述第一節點包括網絡接口處理器(NIP)和高帶寬存儲器(HBM)，並且其中所述HBM的帶寬至少為一千兆字節。

【請求項41】一種用於在基於乙太網的網絡中傳輸分組的第一節點，所述第一節點包括：

一個或多個處理器，包括：

先進先出(FIFO)存儲器，其被配置為存儲與多個鏈路相關聯的時序和狀態信息，其中所述第一節點被配置為使用乙太網協議在所述多個鏈路上向一個或多個其他節點傳輸分組；

定時器，其被配置為根據時間段滴答，其中所述定時器與所述多個鏈路相關聯；和

邏輯電路，其被配置為：

基於定時器上的相應滴答來訪問FIFO存儲器的條目，和

基於與所述多個鏈路中的第一鏈路相關聯的時序和狀態信息，確定重放與第一鏈路相關聯的至少一個分組，

其中所述乙太網協議是有損的。

【請求項42】根據請求項41所述的第一節點，其中所述邏輯電路被配置為以輪詢方式訪問FIFO存儲器的條目。

【請求項43】根據請求項41所述的第一節點，其中所述定時器被配置為基於與FIFO存儲器的條目相關聯的活動鏈路的數量來調整時間段，其中所述活動鏈路被包括在所述多個鏈路中。

【請求項44】根據請求項41所述的第一節點，其中所述邏輯電路被配置為基於與所述多個鏈路中的第二鏈路相關聯的時序和狀態信息來確定退出與第二鏈路相關聯的分組。

【請求項45】根據請求項44所述的第一節點，其中與第二鏈路相關聯的分組存儲在第一節點的本地存儲裝置中，並且其中所述邏輯電路使本地存儲裝置響應於確定退出與第二鏈路相關聯的分組而丟棄與第二鏈路相關聯的分組。

【請求項46】根據請求項41所述的第一節點，其中所述邏輯電路被配置為基於與所述多個鏈路中的第二鏈路相

關聯的時序和狀態信息來確定關閉第二鏈路。

【請求項47】根據請求項41所述的第一節點，其中與所述多個鏈路中的第一鏈路相關聯的時序和狀態信息指示第一節點在用於重放分組的閾值持續時間內沒有接收到對接收到與第一鏈路相關聯的至少一個分組的確認。

【請求項48】一種用於基於乙太網的通信的第一節點，所述第一節點包括：

一個或多個處理器，其被配置為實現傳輸層僅硬件乙太網協議，

其中所述傳輸層僅硬件乙太網協議是有損的，並且

其中所述一個或多個處理器包括硬件鏈路定時器，所述硬件鏈路定時器被配置為確定在傳輸層僅硬件乙太網協議下傳輸的分組以進行重放。

【請求項49】根據請求項48所述的第一節點，其中所述第一節點根據傳輸層僅硬件乙太網協議在第一鏈路上傳輸第一多個分組並在第二鏈路上傳輸第二多個分組，並且其中所述硬件鏈路定時器包括：

先進先出(FIFO)存儲器，其被配置為將與第一鏈路相關聯的時序和狀態信息存儲在FIFO存儲器的第一條目中，並將與第二鏈路相關聯的時序和狀態信息存儲在FIFO存儲器的第二條目中。

【請求項50】根據請求項49所述的第一節點，其中所述硬件鏈路定時器包括與根據時間段滴答的多個鏈路相關聯的定時器，其中所述硬件鏈路定時器以定時器的輪詢方

式滴答來訪問FIFO存儲器的條目，其中所述條目包括第一條目和第二條目。

【請求項51】根據請求項50所述的第一節點，其中所述硬件鏈路定時器被配置為基於與FIFO存儲器的條目相關聯的活動鏈路的數量來調整時間段，並且其中所述活動鏈路包括第一鏈路和第二鏈路。

【請求項52】根據請求項50所述的第一節點，其中所述硬件鏈路定時器被配置為：

基於存儲在FIFO存儲器的第一條目中的與第一鏈路相關聯的時序和狀態信息，確定重放第一多個分組中的至少一些；和

基於存儲在FIFO存儲器的第二條目中的與第二鏈路相關聯的時序和狀態信息，確定退出第二多個分組。

【請求項53】根據請求項52所述的第一節點，其中所述第二多個分組存儲在第一節點的本地存儲裝置中，並且其中所述硬件鏈路定時器使本地存儲裝置響應於確定退出第二多個分組而丟棄第二多個分組。

【請求項54】根據請求項52所述的第一節點，其中與第一鏈路相關聯的時序和狀態信息指示第一節點在用於重放分組的閾值持續時間內沒有接收到對接收第一多個分組之一的確認。

【請求項55】一種在基於乙太網的網絡中的第一節點處實現的計算機實現的方法，所述計算機實現的方法包括：

將與多個鏈路相關聯的時序和狀態信息存儲在第一節點的先進先出(FIFO)存儲器中，其中所述第一節點被配置為使用乙太網協議在所述多個鏈路上向一個或多個其他節點傳輸分組；

基於硬件定時器的相應滴答來訪問FIFO存儲器的條目；和

基於與所述多個鏈路中的第一鏈路相關聯的時序和狀態信息，確定重放與第一鏈路相關聯的至少一個分組，

其中所述乙太網協議是有損的。

【請求項56】根據請求項55所述的計算機實現的方法，其中以輪詢方式訪問FIFO存儲器的條目。

【請求項57】根據請求項55所述的計算機實現的方法，進一步包括：

基於與FIFO存儲器的條目相關聯的活動鏈路的數量來調整硬件定時器的時間段，其中所述活動鏈路被包括在所述多個鏈路中。

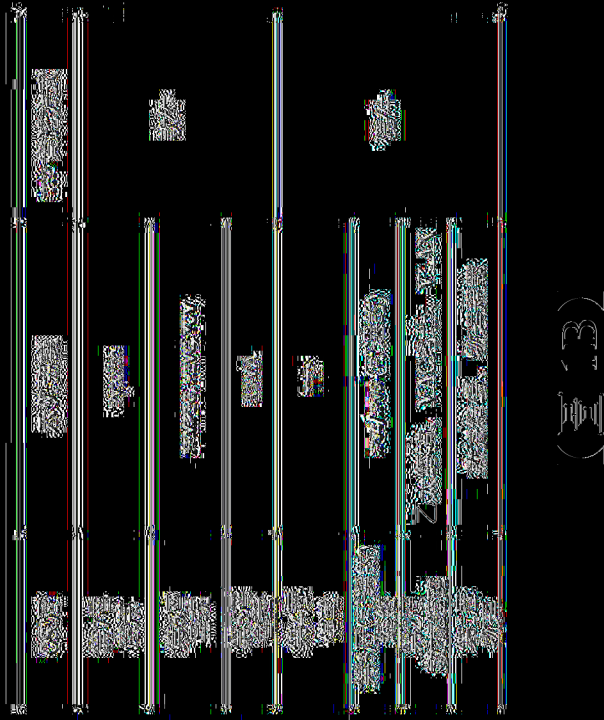
【請求項58】根據請求項55所述的計算機實現的方法，進一步包括：

基於與所述多個鏈路中的第二鏈路相關聯的時序和狀態信息，確定退出與第二鏈路相關聯的分組。

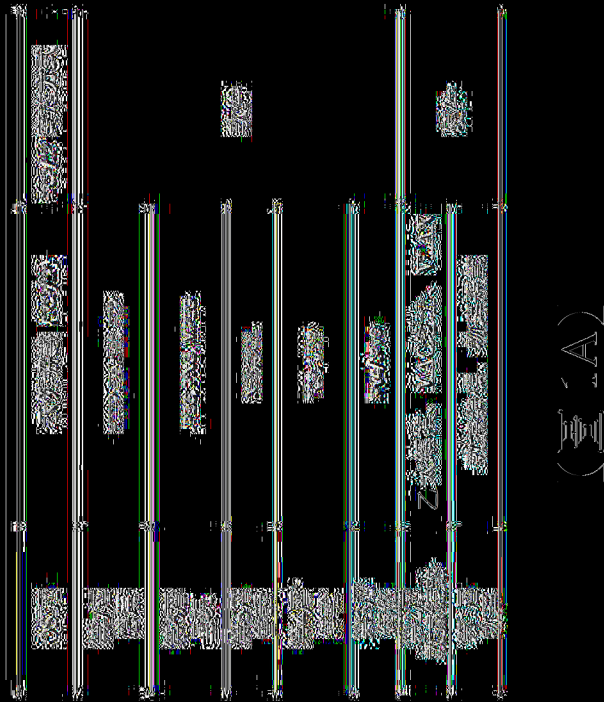
【請求項59】根據請求項55所述的計算機實現的方法，進一步包括使與第一鏈路相關聯的至少一個分組被重放。

【請求項60】根據請求項55所述的計算機實現的方

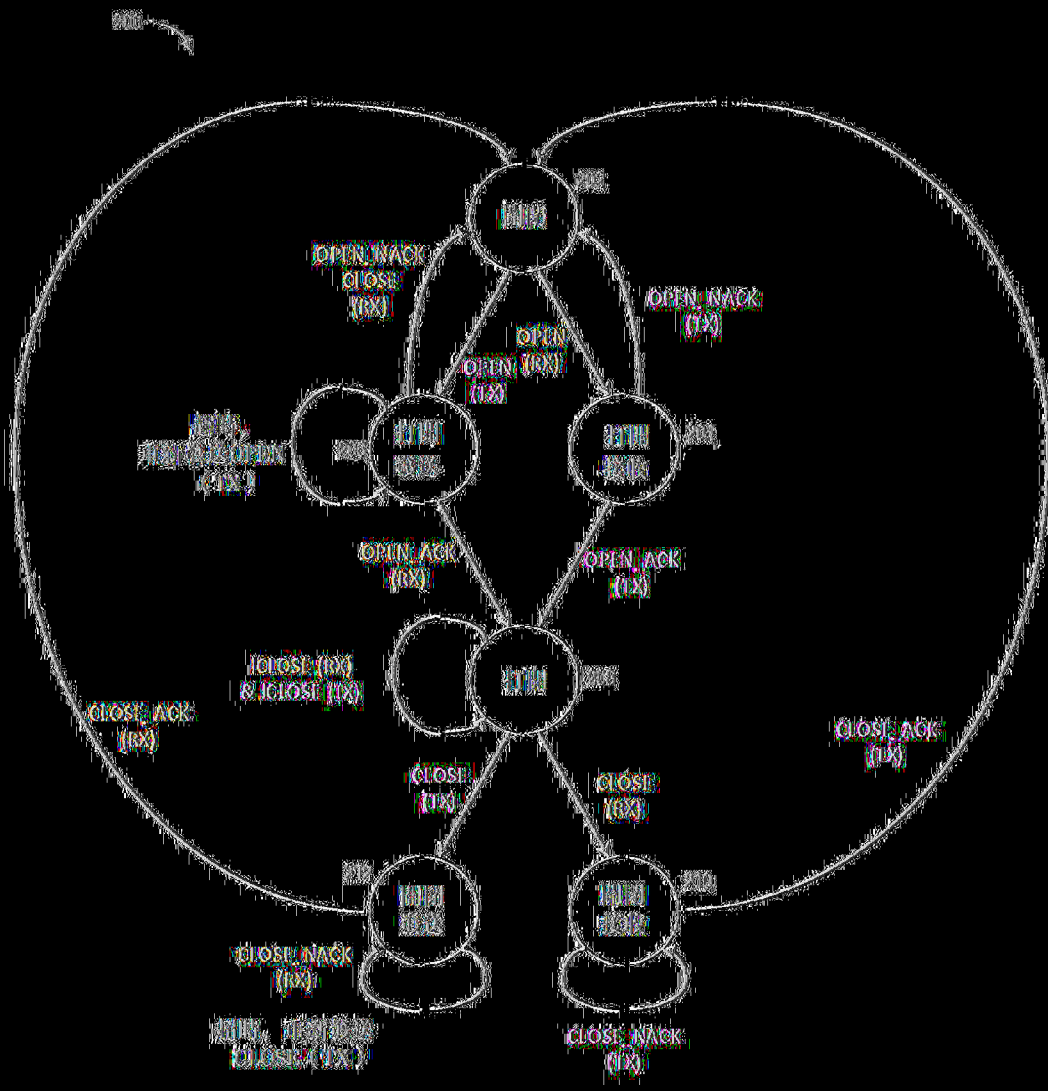
法，其中與所述多個鏈路中的第一鏈路相關聯的時序和狀態信息指示第一節點在用於重放分組的閾值持續時間內沒有接收到對接收到與第一鏈路相關聯的至少一個分組的確認。



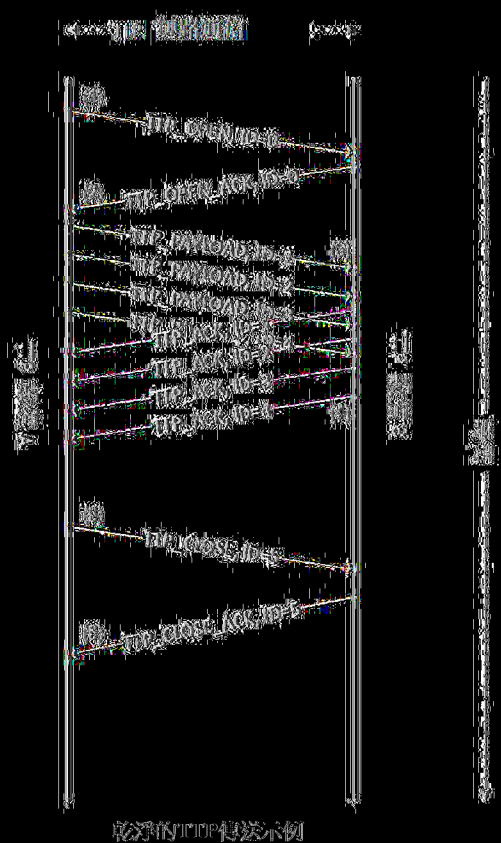
(圖 13)



(圖 14)

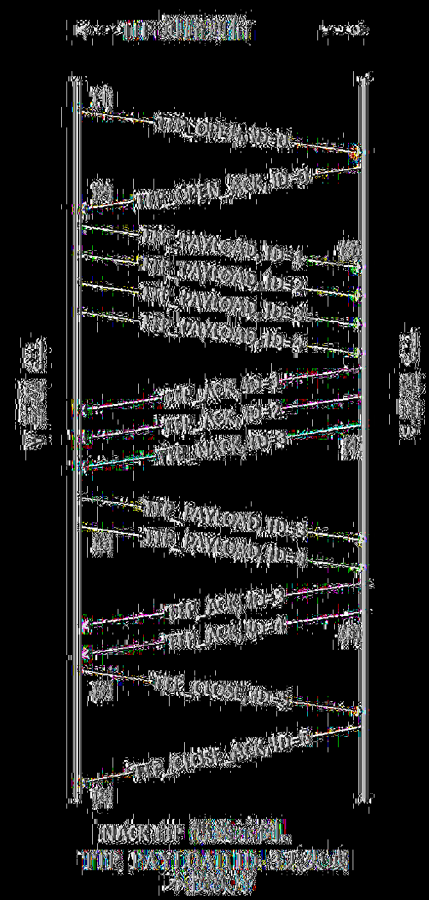


$$(|\{ \begin{smallmatrix} 1 & 2 \\ 1 & 1 \end{smallmatrix} \}| 2)$$



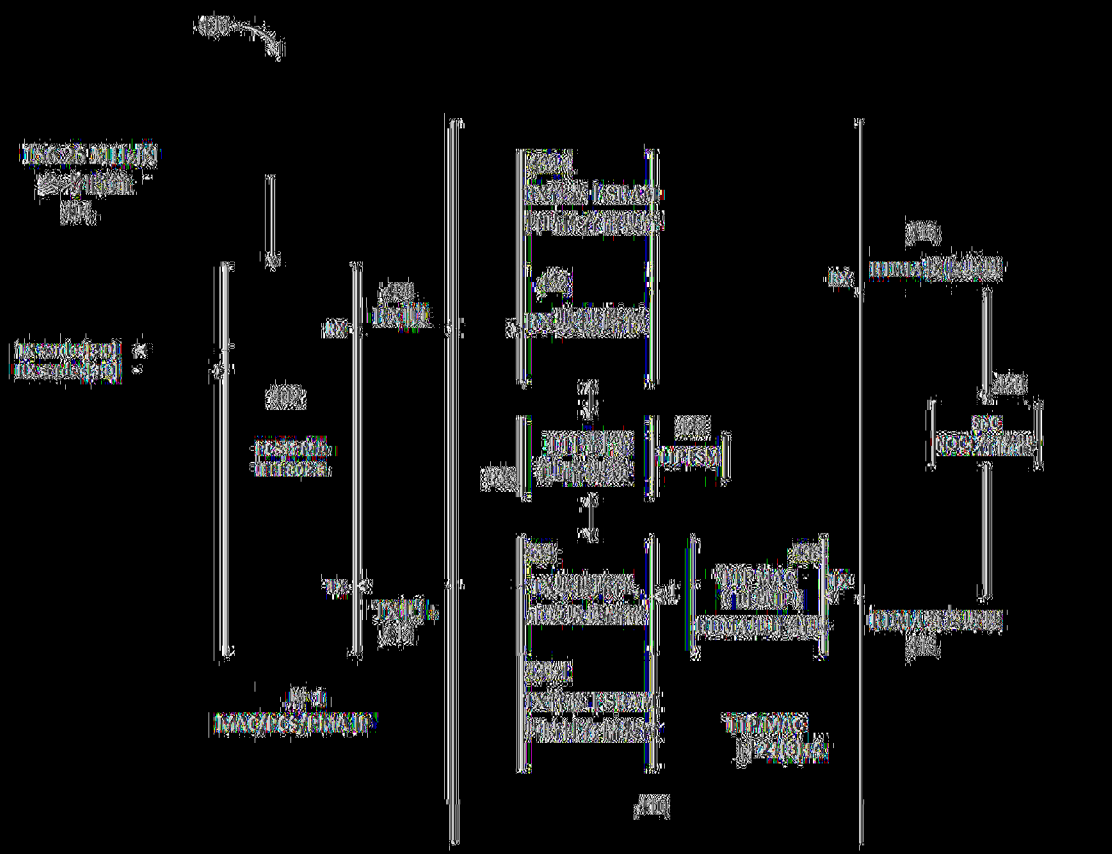
乾對的TDP傳送示例

(3A)

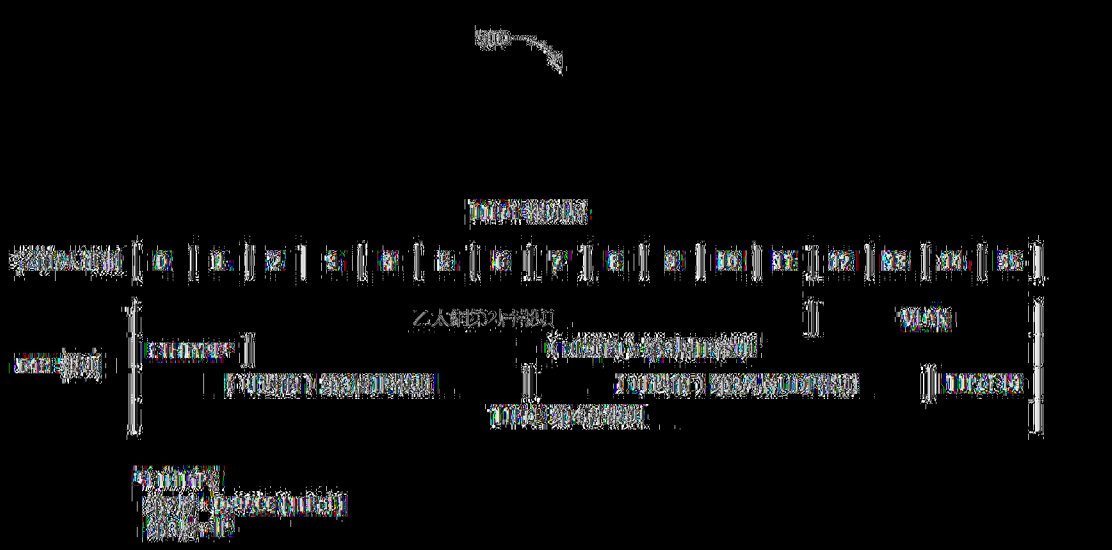


乾對的TDP傳送示例

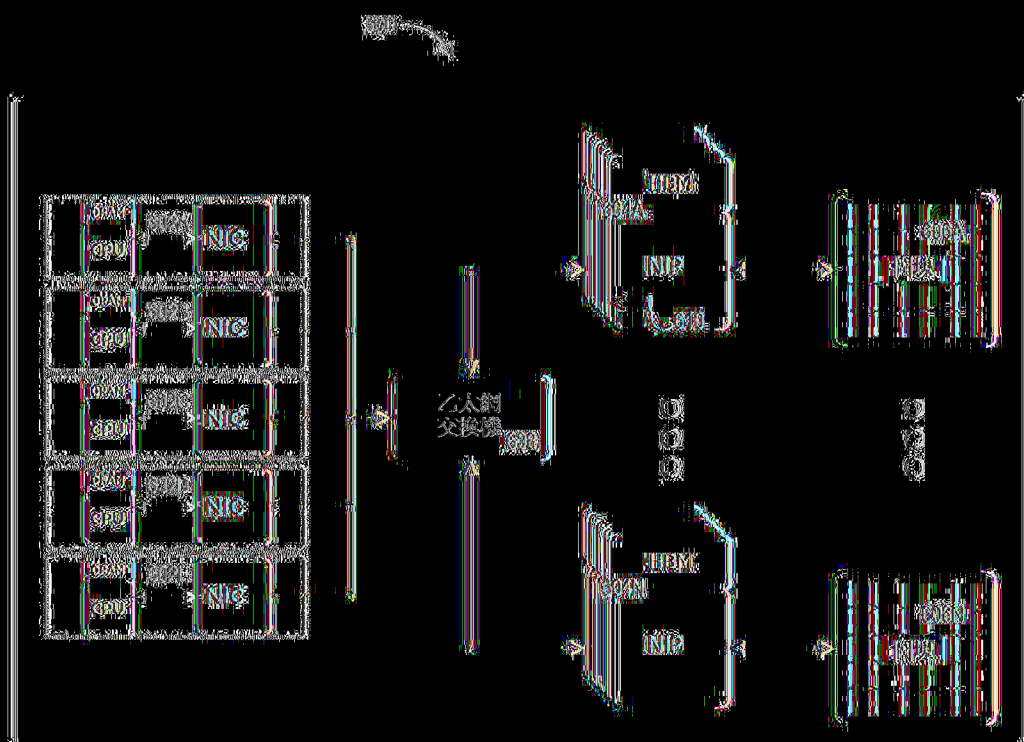
(3B)



$$\left| \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) 4 \right|$$



$$\left| \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) 5 \right|$$

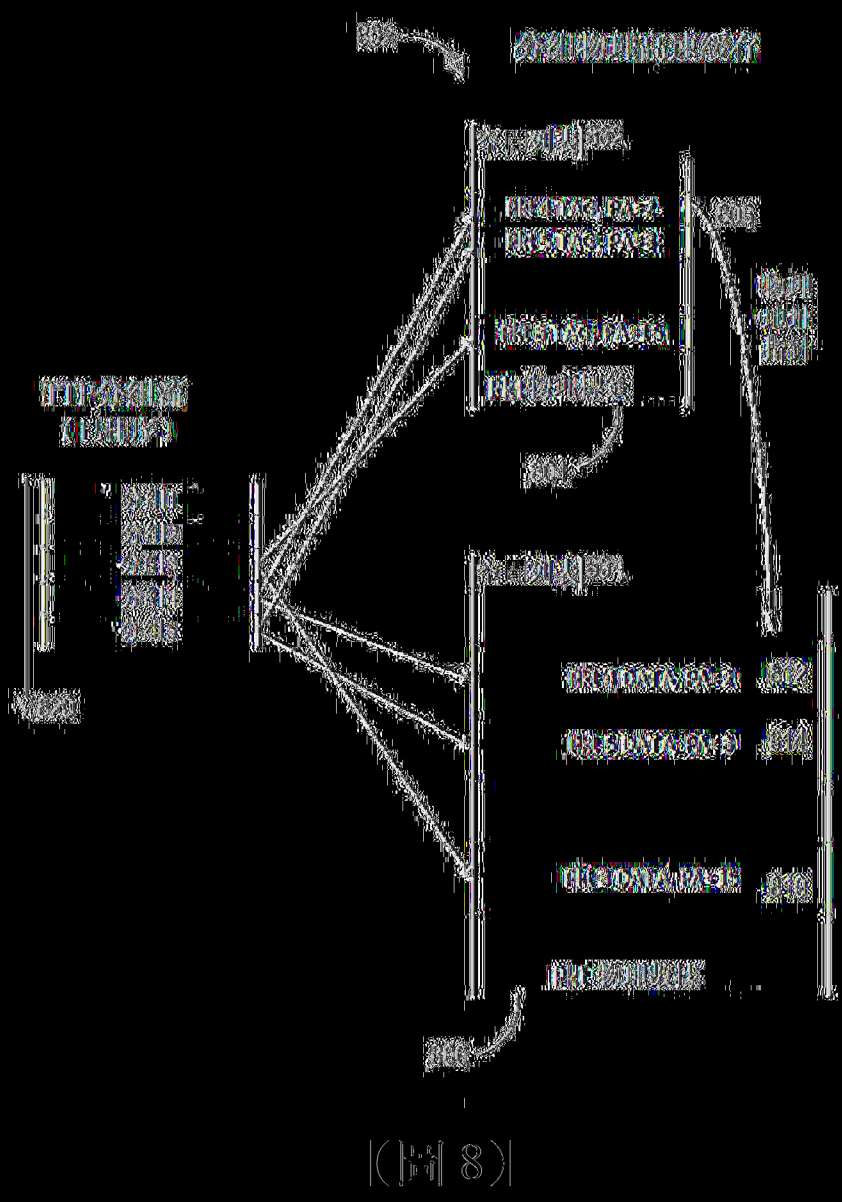


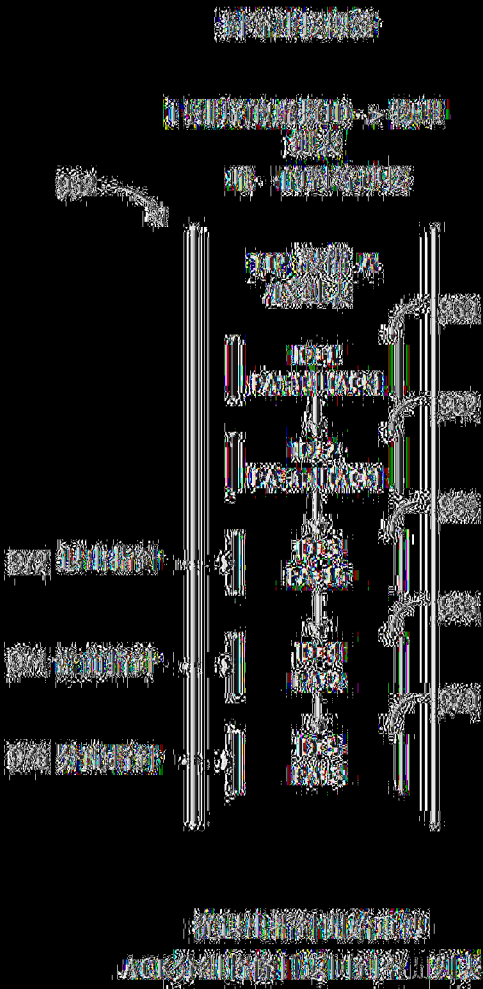
(图 6)

[illegible]

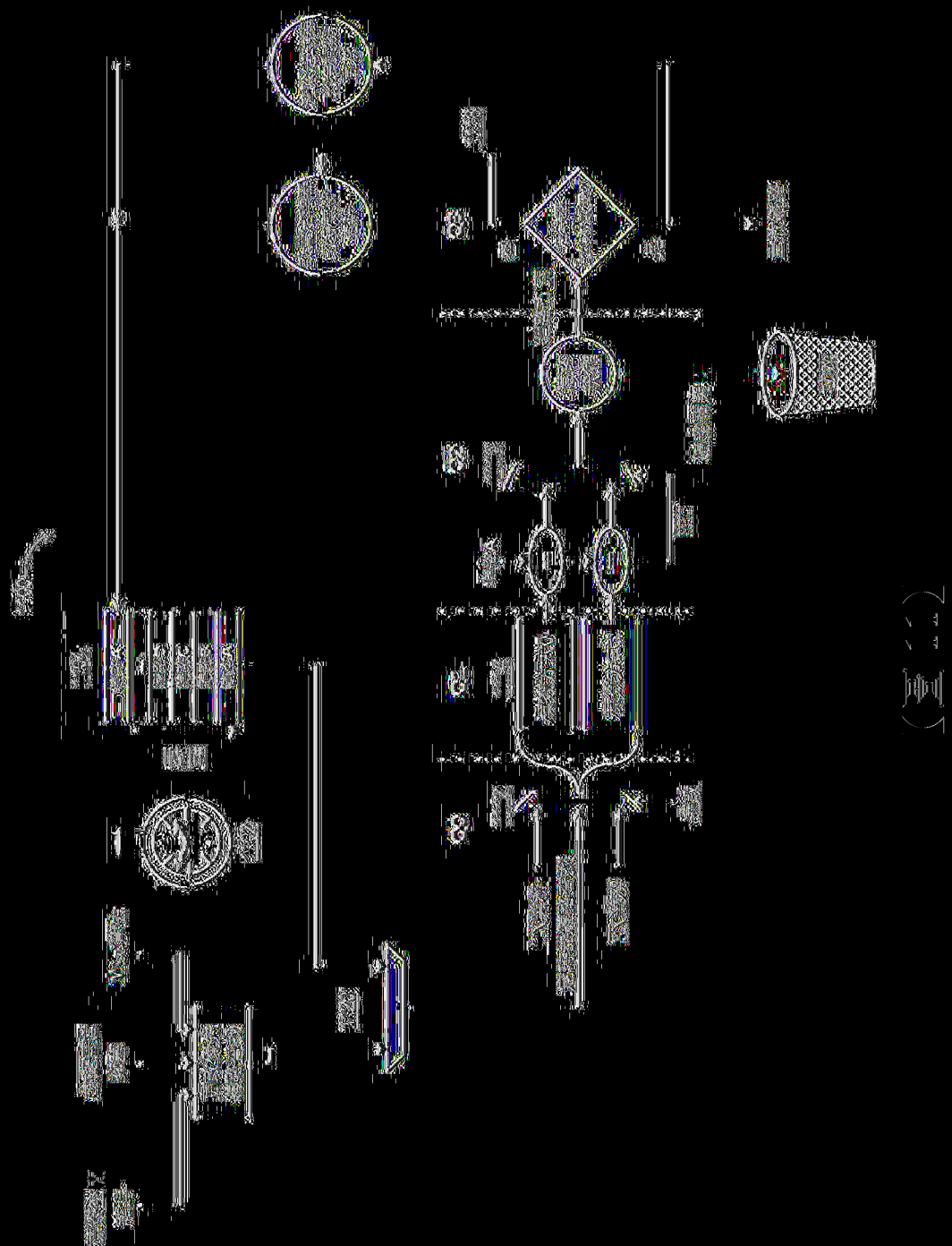
$$\left| \left(\begin{bmatrix} \mathbb{E} u \\ \mathbb{E} v \end{bmatrix} \otimes \mathbb{A} \right) \right|$$

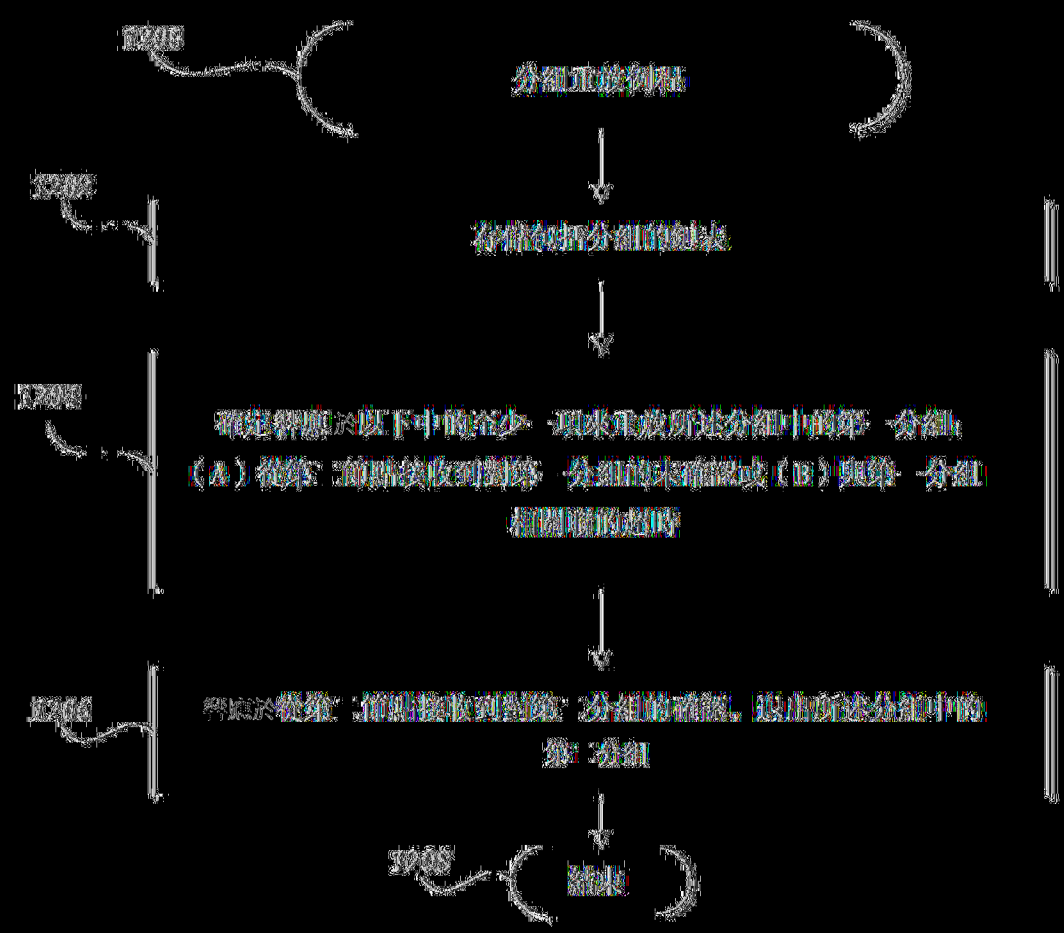
[illegible]



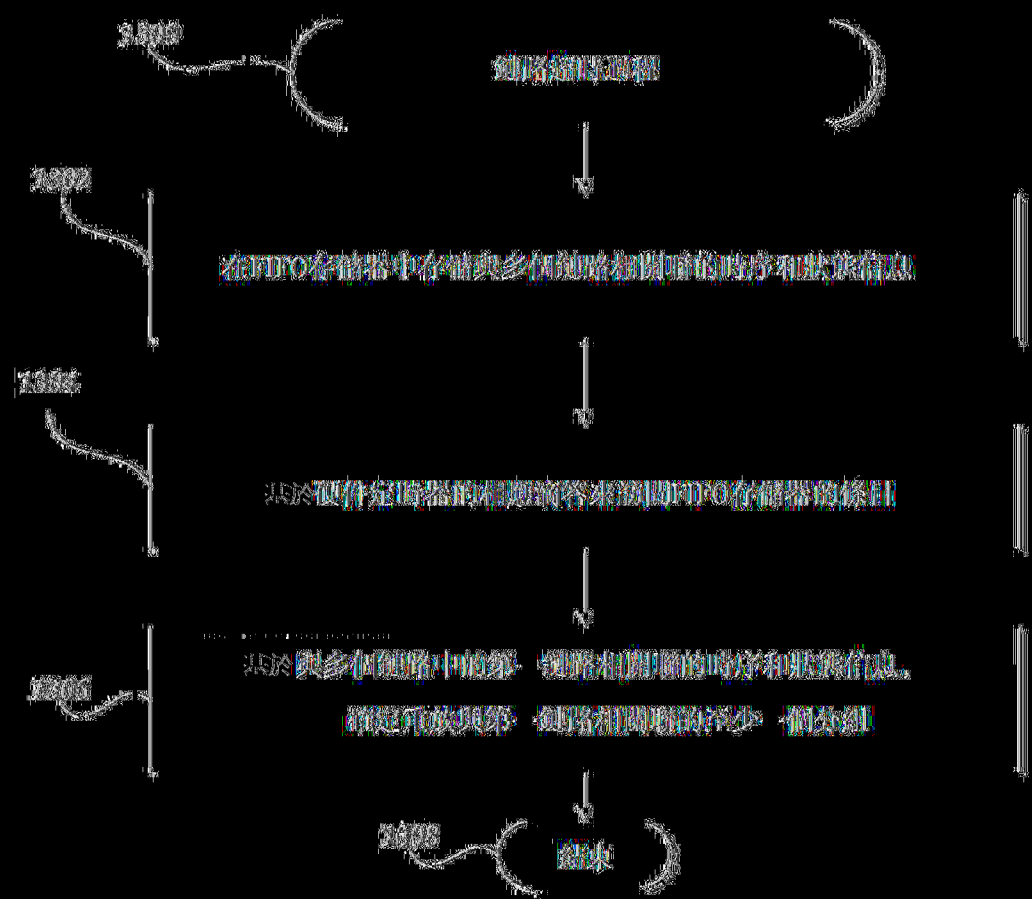


$$\left| \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) 9 \right|$$





|(圖 12)|



[(13)]