

WIND RIVER

以多核心技術來最佳化網路性能

提要

在網路通信設備的評測中，性能和成本一直是最關鍵的要素。然而，性能包括了多個方面，包括吞吐能力、時延和CPU佔用率等。即便對於容量小於1GB的系統，可預測的回應時間和可用于運行應用的CPU週期都至關重要。多核心晶片的出現為性能的提升和降低成本帶來了機遇。在多個內核之間高效率地分佈網路通信功能，系統就可以實現比前一代產品更高的吞吐能力、更低的CPU佔用率、更小的尺寸和更低的成本。本文描述了作業系統、協議棧和多核心晶片的有效集成將會給網路通信行業帶來怎樣的變革。

更高性能帶來的挑戰

根據摩爾定理，處理器的能力每兩年就要翻一倍^[1]。由此對所有使用微處理器的設備產生了廣泛的影響，其中當然包括網路通信設備。性能更強大的終端節點能夠更快地處理資料，這也是網路通信頻寬需求不斷增長的動力。在過去的15年內，局域網（LAN）的傳輸速率已經增長了1000多倍。

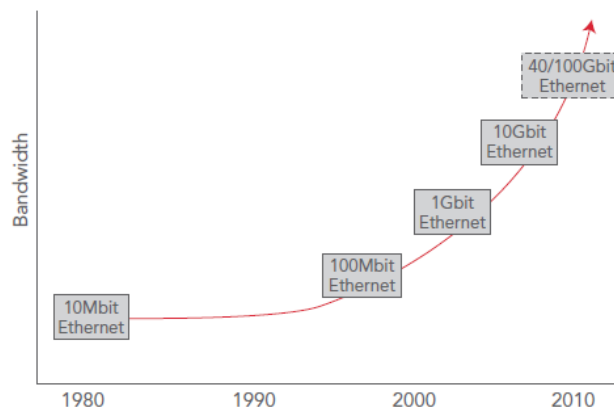


圖1：局域網資料速率的演進

廣域網路（WAN）的資料速率雖然沒有局域網速率那麼快，但是也呈現出指數級的增長。

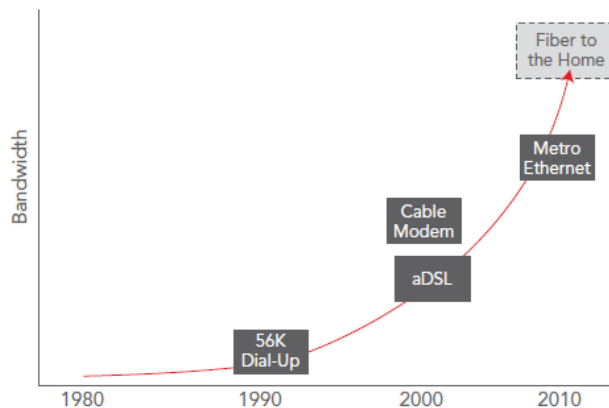


圖2：廣域網路資料速率演進

發展最快的網路通信技術是無線通訊。由於不再受到線纜連接的牽絆，無線通訊為用戶互連提供了極大的便利性。

有人爭辯說，多媒體內容的需求已經存在多時，早就在等待相應的網路通信技術來支援了。視頻和音訊檔不僅體積比純文字資料檔案大得多，而且對時間延遲更為敏感。

語音、視頻和資料的融合需要更精密的通信設備以滿足低時延的需求。如今的家庭網關需要實現Internet接入、VoIP語音通信和流媒體等多種業務混合一體的處理能力。

同樣，像蘋果iPhone這類手機設備中也融合了語音、資料、音樂、Internet和多媒體等多種功能，而且把這些功能放在了更小的設備中。

所有這些發展趨勢都需要高頻寬、低時延的網路通信技術來實現，不論是對於終端使用者設備，而且包括各種接入、彙聚和核心部件。設備製造商面臨的挑戰是在開發週期縮短、產品利潤空間壓縮的壓力下，為市場提供更高性能的平臺。解決所有這些需求和挑戰需要全新的解決方案。多核心網路就是解決方案。

多核性能

在過去幾十年裡，處理器能力每兩年就翻一倍。然而，近幾年處理器速率上升曲線開始變得平緩，這是由於受到發熱和功耗等因素限制，無法再通過增加電晶體數量來提升處理器性能。

但是，多核心處理提供了新思路。通過併發地使用多個內核心，處理性能可以進一步提升以滿足高性能的需求。

全球領先的各大處理器晶片廠商都開始推出多核心晶片，在單個晶片內集成了多個處理內核心。通過非常快速的任務間資料交換，虛擬內核心或執行緒可以進一步細分內核心資源。

多核心處理器晶片的性能依據時脈速率和內核數量而不同。目前已經有16—32個內核的處理器晶片。這些晶片中大多數都集成了網路處理功能，減小了由傳統網路通訊協定軟體所造成的時延。

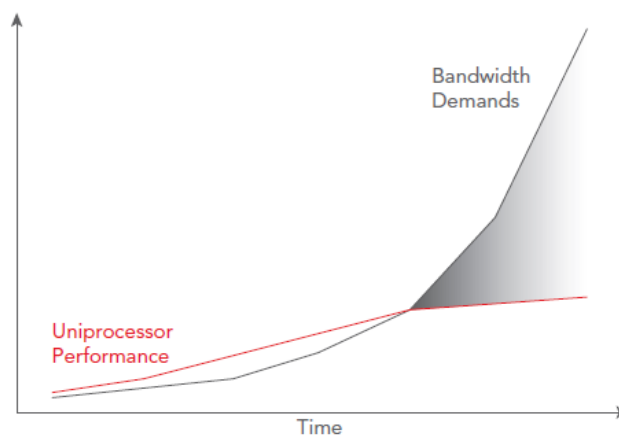


圖3：單個處理器性能差距

多核心處理方法

多核心軟體可以由多種模式實現^[2]。在採用對稱式多處理結構（SMP）模式實現的系統中，在運作系統和任務時，多個內核心基本上是可以互換的。有一種SMP採用了聯姻（affinity）或CPU預留技術來指定任務與某個內核心的綁定，由此使其變成較為高效的專用處理器。

非對稱式處理結構（AMP）通常是指運行著多個作業系統的架構。Supervised AMP採用了虛擬化技術對各種處理單元進行抽象，例如記憶體、內核心或設備等。

為了發揮新型晶片的優勢，必需設計出新的軟體。一種常見的誤解是，為單核心處理器環境編寫的軟體在多核心處理器環境下自然能夠運行得更快更好。讓我們以機器人為例，在裝配生產線上經常使用機器人手臂來搬動箱子。當採用單處理器運行時，每分鐘能夠搬運12個箱子。如果同樣的系統和軟體以SMP模式採用多核心處理器運行的話，機器人手臂並不會運行得更快，每分鐘仍然只能搬運12個箱子。但是，如果將軟體面向多核心處理器技術進行重新編寫，系統就能夠使用更多的處理能力去執行其他的任務。例如，如果在上述機器人控制多核系統中，將第二個處理器用於控制另一個機器人手臂，並且與第一個手臂交叉配合，可以實現每分鐘搬運24個箱子，使生產效率加倍。此外，第二個處理器還可以用於控制傳送帶、遙感探測回饋或分擔第一個處理器上的任務負載，將生產效率提高到每分鐘搬運15個箱子。從這個例子中顯而易見，僅僅在硬體方面向多核心處理技術的轉變不會自動地提升性能，必需通過內核心間的相互配合和交互，才能將充分發揮多核心處理器的性能。

以多核心處理網路通訊協定

網路通信協議棧中最常見的協議就是IP和TCP。這些協定在互聯網技術中廣泛使用，實際上還用於所有使用網路連接功能的行業和產品。同樣，網路通信設備也必須支援一套通用的網路通訊協定，由此支持保持所有行業實現互連的基礎架構。傳統方式下，這些協定被當做統一處理資料包的單一棧。

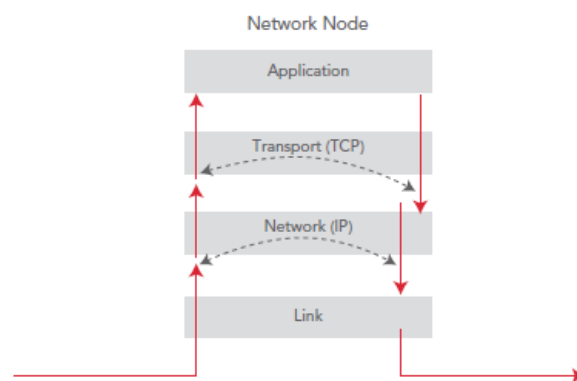


圖4：傳統的單一棧資料包處理

資料包進入單一棧後的處理步驟可以想像為一個過程狀態機。資料包首先被物理鏈路層（通常是乙太網）界面接收，然後排隊進入網路層（IP層）。IP層負責確定資料包目的地是本機還是需要被繼續轉發。此外，作為IPsec或IKE等安全協議的一部分，IP層還需要完成對應的資料包加密操作。如果資料包的目的地就是本機，那麼它將跳過本層而被轉到上一層

協議（通常是TCP或UDP）進行處理。更多的安全功能需要通過安全通訊端層（SSL）進行處理。如果資料封包還是需要送往本機，那麼它將被轉到更高的應用層協議，包括FTP、SMTP、Telnet和HTTP等。在單核心系統甚至純SMP系統中，所有這些網路處理過程都需要競爭處理器工作週期。

單一的分散式網路通訊協定棧架構並不能完全反應所有的應用場景。例如，同樣是不斷重複執行的步驟，當用於數據包轉發時和用於數據包就是不同的。交換機和閘道在第三層存在區別，交換機負責在介面和web伺服器間轉發數據包，而閘道負責接收（或終止）資料請求並返回HTML資料的頁面。對網路通訊協定棧的優化需要針對不同的應用場景採用不同的架構。

另外，系統設計的選擇和確定必須根據協定棧的哪一層進行分發。如果大多數的連接需要在第四層（TCP）進行管理，最好採用能夠跨處理器內核心分配多TCP實例的架構。然而，多實例必然帶來協定複雜度的增加，需要內核心間更多的交互，從而導致時延的增加和記憶體頻寬的受限。這個問題在頻寬為1G或2G的情況下可能還不明顯，但隨著系統吞吐頻寬的進一步擴展，它將成為嚴重的制約因素之一。成本/收益分析必須綜合考慮整個系統的目標

在交換機、路由器和閘道等高性能網路通信設備中，絕大多數網路數據包的轉發都發生在IP層。作為一種高頻次重複的任務，數據包轉發功能尤其適於採用多核心技術來實現。首先，需要將IP協定中實現轉發功能的代碼與地址建立邏輯代碼進行分離。當接收數據包中的地址是第一次出現時，將地址記錄在地址資料表格中。由於此項任務只需執行一次，資料表的維護可以採用多用途內核心中的“慢速通道（slow path）”，它的功能就像是作業系統中的傳統網路通訊協議棧處理器。“快速通道（fast path）”處理器僅僅需要檢測接收數據包的目的地地址，並且查找其緩存資料表。如果地址被找到，快速通道處理器快速確定對應的輸出埠，並將數據包及時送入轉發隊列中。

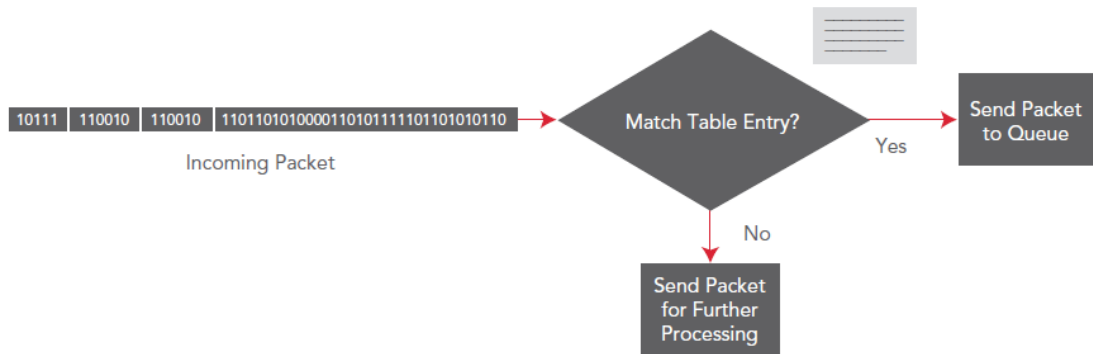


圖5：數據包處理邏輯狀態

通過這種方式實現的IP資料包分類演算法適於各種高效的第三層網路功能，包括數據包轉發、網路地址翻譯（NAT）、存取控制清單、IPsec和其他加密資料功能等。某些處理器晶片具有針對資料加密的專用引擎，可以與CPU併發運行。

雖然用於實現這種快速通道式轉發的代碼相對更簡單，但高達數G比特的介面產生的高資料流程速率還是會對處理器內核心帶來巨大的壓力。採用更多的轉發核心可以部分地緩解高輸送量帶來的問題，但必需結合採用其他的系統設計技術，才能避免系統性能不會因為這些瓶頸而達到上限。

多核心設計考慮的因素

基於多內核心設計的分散式網路通訊協議棧能夠極大地提升網路通信設備的系統性能，但必須考慮好相關因素。當多個內核心共用記憶體時，某一時刻只有一個內核心能夠更新資訊，其他內核心必須等待訪問權。這種互鎖安全機制一般通過軟體使用semaphores、spinlock和mutual exclusion等技術實現。由於多核心處理器必須相互等待才能完成共用記憶體中的資料結構修改，隨著內核心數量的增加，也會帶來更多的衝突。這就是系統輸送量性能曲線在某個時間點出現下拐或趨平時所呈現的情況。

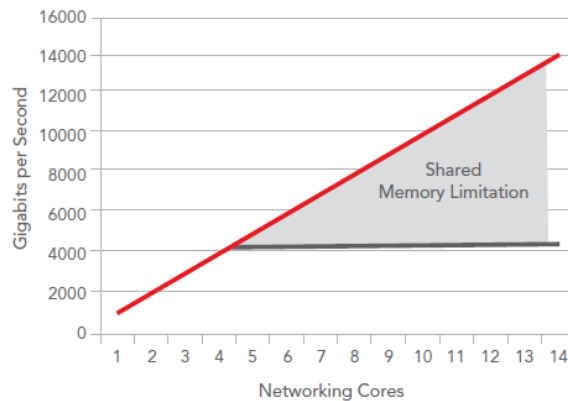


圖6：共用記憶體帶來的頻寬限制

必須對多核心架構代碼設計給予仔細考慮，才能將互鎖帶來的影響降到最小。如果可能，應該儘量應用晶片中無互鎖代碼的功能特性，以避免軟體性能極限。

緩存（Caching）與散列（Hashing）技術

幾十年來，指令緩存技術（Instruction caching）不斷被研究和改善，用於實現更高效的資料處理。高頻次重複的任務很適合採用多核心網路加速技術，但是如果指令緩存沒有得到最佳化的配置，也很容易造成性能的反常下降。讓我們設想一個每秒能夠處理140萬個數據包的IP轉發核心。單單一個緩存的缺失就會導致處理迴圈需要多花費15%的時鐘週期，從而造成處理能力的相應下降。調校代碼以匹配指令緩存，這項工作是非常精細繁雜的，但是高效地使用緩存技術能夠為網路通訊協定處理的性能帶來極大地提升。

同樣，我們也需要高效地處理各種資料表結構，包括埠映射、位址翻譯、流量數據和安全關聯等，從而避免性能曲線的下拐。散列（Hashing）技術是存取這些數據的高效手段，但是其前提是必須分配足夠的空間以避免衝突發生。如果兩個數據表入口都控制相同的散列值，那麼數據訪問效率將會降低。因此，為了應付預期的數據流程速率，系統必須預先就進行正確地設計和配置，才能避免效率的下降。

多核心性能獲益

在網路通信協議棧中應用多核心可以從兩方面獲益。一方面，也是最明顯的方面，是提升吞吐能力。吞吐能力的度量單位通常是“包/秒”或者“兆比特/秒”。通常，我們常誤認為千

兆乙太網（Gigabit Ethernet）能夠完全支持每秒1千兆比特的資料輸送量。實際上，一部分頻寬會被數據包間隙或頭部開銷等所佔據。

封包頭： 8位元組
 數據包間隙： 12位元組/20位元組
 20位元組

在採用64位元組乙太網幀的1Gb/秒鏈路中，線上路上實際發送的資料為20+64=84位元組。也就是說， $20/84 = 23.8\%$ 的線路容量（即238Mbps）被用於頭部開銷，剩下76.2%（即762Mbps）用於傳送資料。當資料幀大小增加時，它們的發送次數可以降低，從而開銷佔據頻寬的比例也隨之下降。下表顯示了各種幀大小情況下的最大畫面播放速率和可用資料量。

Frame Size	Throughput	FPS
64	762Mbps	1,488,000
128	865Mbps	844,595
256	928Mbps	452,900
512	962Mbps	234,962
1024	981Mbps	119,732
1280	985Mbps	96,154
1518	987Mbps	81,274

表1：以Mbps和幀、每秒(FPS)衡量1Gb乙太網理論最大有效載荷

常用的測量基準還有數據包轉發（指一個介面接收到資料包並轉發至另一介面）和數據包終止（指數據包被處理並終止）。根據被測試網路設備的類型，某些測試手段可能更適合於測量設備性能。

圖7展示了通過基於Wind River VxWorks 6.7平臺的500MHz Cavium OCTEON 3860測試的IPv4數據包轉發速率。

Frame Size	Traditional IP Forwarding	Multicore Network Acceleration	Percent Increase
64	18	762	4133%
128	34	865	2444%
256	64	928	1350%
512	127	962	657%
1024	277	981	254%
1280	343	985	187%
1518	410	987	141%

Measured with VxWorks 6.7 EAR release on Cavium 3860, 500MHz, two cores/
 one forwarder; VxWorks core is 99% to 100% idle during test

圖7：採用多核心加速的網路數據包處理

除了網路吞吐能力以外，系統的整體性能還依賴於可供其他系統任務使用的處理能力。如果所有的資源都被耗在數據包處理方面，那麼即使最簡單的系統管理任務也可能導致資源耗盡和性能下降。也就是說，就算這些設備在測試過程中能夠獲得期望的吞吐速率，但前文所描述的瓶頸在也會非常明顯地表現出來。為了讓網路負載分流真正發揮明顯的效益，必須讓負載分流演算法對核心作業系統的依賴性降低到最小的程度。

結論

多核心晶片提供了實現快速數據包處理的全新方案。通過在單個晶片上整合多個處理內核心，網路設備可以被設計得體積更小、功耗更低、成本更低而性能更高。然而，要獲得可擴展的線速性能，必需細緻地考慮硬體和軟體架構。硬體加速特性可以節省寶貴的處理器週期，應當儘量採用。有效地使用緩存隊列和散列資料表，是實現快速路徑性能最大化的關鍵。對稱多處理模式非常有效，但也必需仔細設計，確保網路通信任務以高度並行化的方式執行。採用專用核心來完成網路加速效率提高甚多，達到滿意的線速性能和可擴展性。

注：

1. 摩爾定律宣稱，積體電路中可以置入的電晶體數量將會每兩年翻一倍。這個定律已經被證明適用於性能、容量和成本等許多相關領域。
2. Device Software Optimization for Concurrent and Consecutive Systems, Wind River, <http://windriver.com/whitepapers/>.
3. Achieving Business Goals with Wind River's Multicore Solution, Wind River, <http://windriver.com/whitepapers/>.

Wind River 就在您身邊

北京代表處	北京市朝陽區望京中環南路9號望京大廈B座18層	郵編: 100102	電話: 010-84777100	傳真: 010-64398189
上海代表處	上海市西藏路585號新金橋廣場3-H, I, J室	郵編: 200003	電話: 021-63585586/87/89/90	傳真: 021-63585591
深圳代表處	深圳市福田區車公廟天安數碼時代大廈A座606室	郵編: 518040	電話: 0755-25333408/3418/4508/4518	傳真: 0755-25334318
西安代表處	西安市高新區科技二路68號西安軟體園秦風閣H103	郵編: 710075	電話: 029-87607208	傳真: 029-87607209
成都代表處	成都市武侯區武青南路10號5棟2單元303室	郵編: 610045	電話: 028-87491282	傳真: 028-87491282

關於風河更多內容請訪問: <http://www.windriver.com> Email: inquiries-ap-china@windriver.com

WIND RIVER

© 2007 Wind River Systems, Inc. The Wind River logo is a trademark, and Wind River is a registered trademark of Wind River Systems, Inc. Other marks are the property of their respective owners.