使用 PCIe 交換網結構在多主機系統進行資源部署最佳化

■作者：Vincent Haché
Microchip Technology Inc. 韌體工程技術顧問

越來越多的資料中心和其他高效能計算環境開始使用 GPU，因為 GPU 能夠快速處理深度學習和機器學習應用中所產生的大量資料。不過，就像許多新的數據中心的創新，提高了應用程序的性能，但這項創新也暴露出新的系統瓶頸。在這些應用中，用於提高系統效能的新興架構涉及透過一個 PCIe 結構在多個主機之間共用系統資源。

PCIe 標準 (特別是其基於樹狀結構的傳統層級) 會限制資源分享的實現方式 (和實現程度)。不過，可以實現一種低延遲的高速結構方法，這種方法允許在多個主機之間共用大量 GPU 和 NVMe SSD，同時仍支援標準系統驅動程式。

PCIe 結構方法採用動態分區和多主機單根 (Signal_Root) I/O 虛擬化 (SR-IOV) 共用。各 PCIe 結構之間可直接路由點對點傳輸。這樣便可為點對點傳輸提供最佳路由，減少根連結埠擁塞，並且更有效地平衡 CPU 資源負載。

傳統上，GPU 傳輸必須存取 CPU 的系統記憶體，這會導致端點之間發生記憶體共用的爭奪。當 GPU 使用其共用的記憶體映射資源而不是 CPU 記憶體時，它可以在本地提取資料，無需先通過 CPU 傳遞資料。這避免了轉傳和鏈路所產生的延遲，因而使 GPU 能夠更高效地處理資料。

PCIe 的固有限制

PCIe 主層級是一個樹狀結構，其中的每個域都有一個根聯合體，從該點可擴展到 “ 葉子 ”，這些 “ 葉子 ” 通過交換網和橋接器到達端點。鏈路的嚴格層級和方向性給多主機、多交換網系統帶來了成本高昂的設計要求。

以圖 1 所示的系統為例。要符合 PCIe 的層級，主機 1 必須在交換網 1 中有一個專用的下行埠，該埠連接到交換網 2 中的專用上行埠。它還需要在交換網 2 中有一個專用的下行埠，該埠連接到交換網

圖 1：多主機拓撲

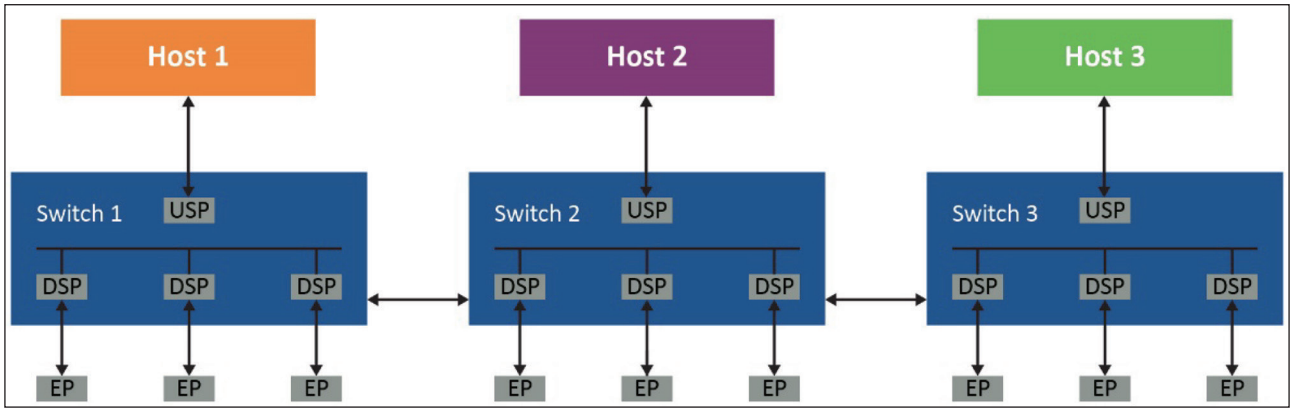
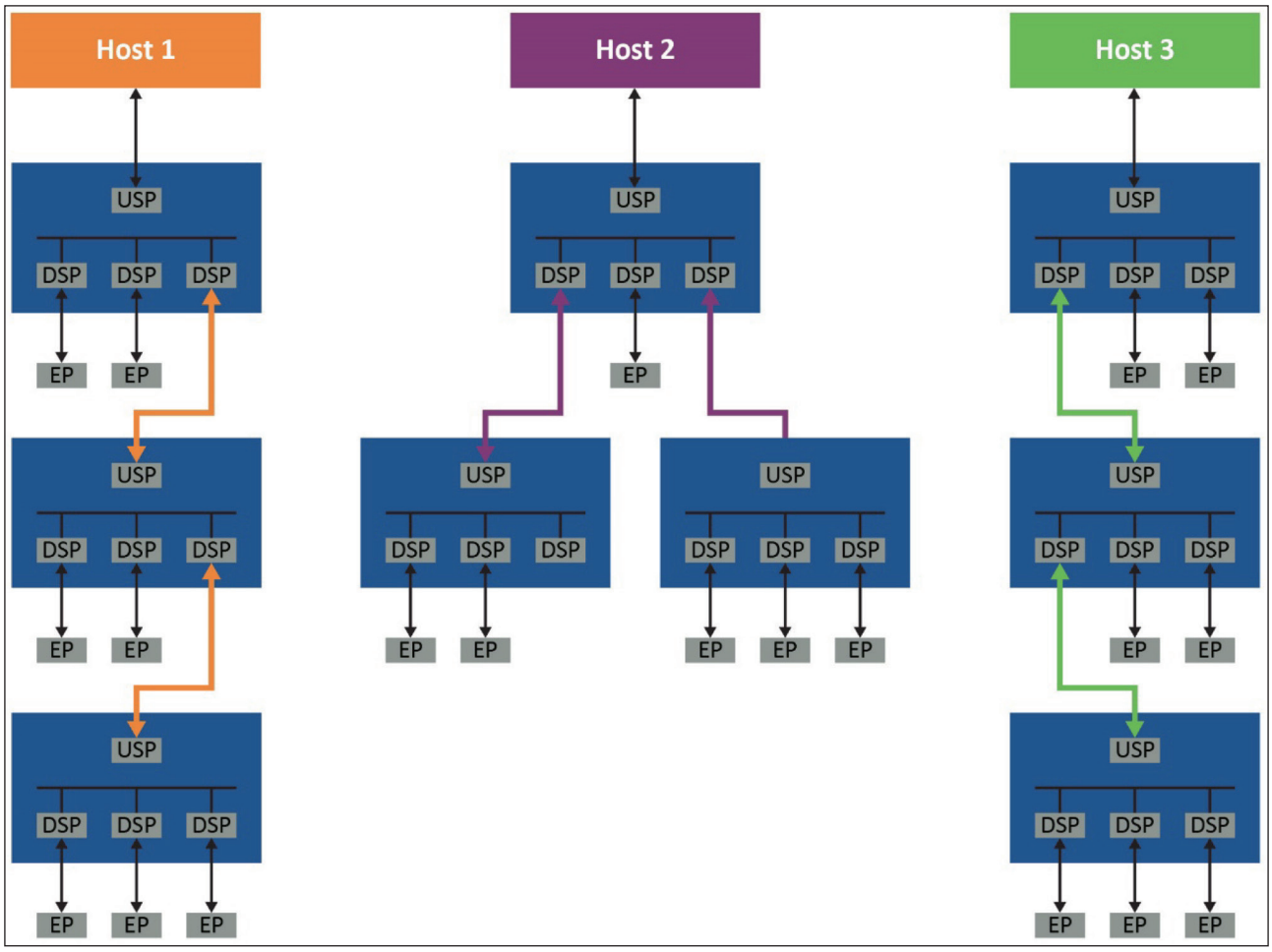


圖 2：每個主機的層級要求



3 中的專用上行埠，依此類推。主機 2 和主機 3 也有類似的要求，如圖 2 所示。

即使是基於 PCIe 樹狀結構的最基本系統，也需要各交換網之間有三個鏈路專用於每個主機的 PCIe 拓撲。而且，由於主機之間無法共用這些鏈路，因此系統會很快變得極其低的效率。

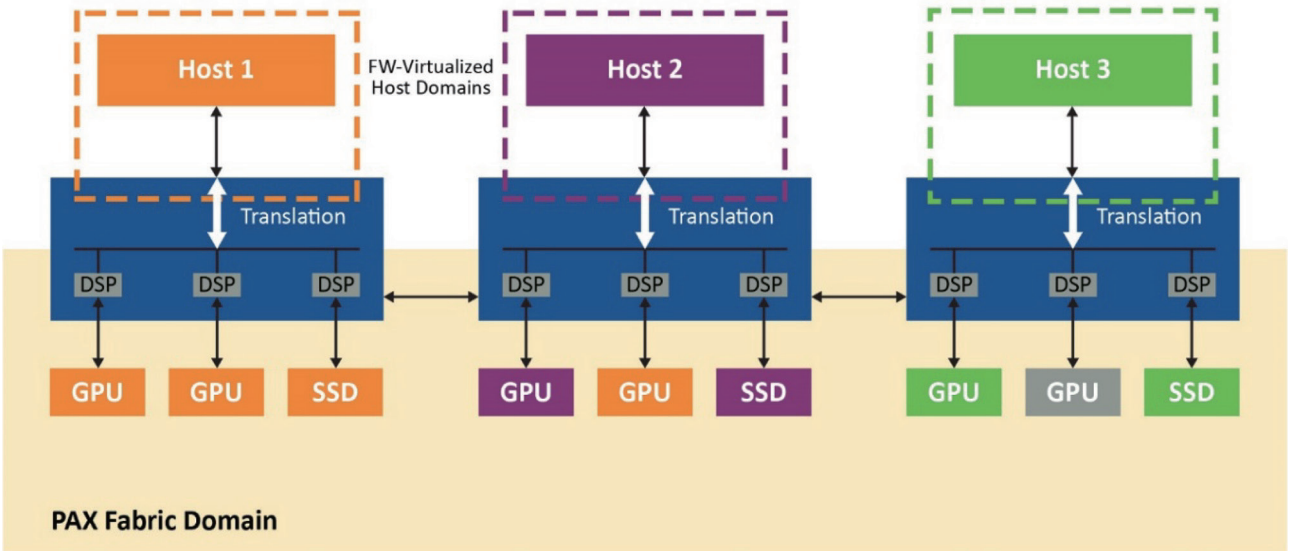
此外，符合 PCIe 的典型層級只有一個根連結埠，而且儘管“多根 I/O 虛擬化和共用”規範中支援多個根，但它會使設計更複雜，並且當前不受主流 CPU 支援。結果會造成未使用的 PCIe 設備 (即端點) 滯留在其分配到的主機中。不難想像，這在採用多個 GPU、存放裝置及其控制器以及交換網的大型系統中會變得多麼低的效率。

例如，如果第一個主機 (主機 1) 已經消耗了所有計算資源，而主機 2 和 3 未充分利用資源，則顯然希望主機 1 存取這些資源。但主機 1 無法這樣做，因為這些資源在它的層級域之外，因此會發生滯留。非透明橋接 (NTB) 是這種問題的一個潛在解決方案，但由於每種類型的共用 PCIe 設備都需要非標準驅動程式和軟體，因此這同樣會使系統變得複雜。更好的方法是使用 PCIe 結構，這種結構允許標準 PCIe 拓撲容納多個可存取每個端點的主機。

實施方法

系統使用一個 PCIe 結構交換網 (本例中為 Microchip Switchtec PAX 系列的成員) 在兩個獨立但可透明交互操作的域中實現：即包含所有端點和結構鏈路的結構域以及每個主機專用的主機域 (圖 3)。主機通過在嵌入式 CPU 上運行的 PAX 交換網韌體保留在單獨的虛擬域中，因此，交換網將始終

圖 3：每個結構的獨立區域

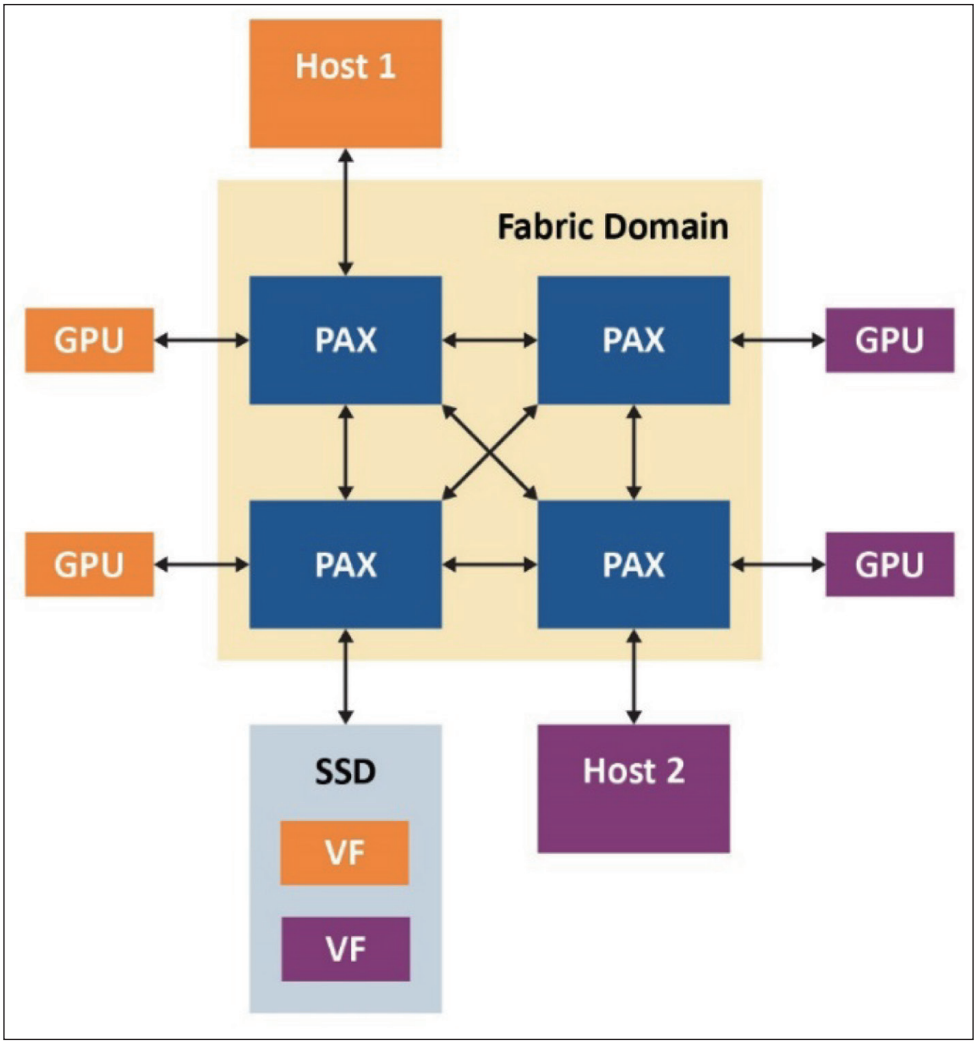


顯示為具有直連端點的標準單層 PCIe 設備，而與這些端點出現在結構中的位置無關。

來自主機域的事務會在結構域中轉換為 ID 和位址，反之，結構域中通訊的非分層路由也是如此。這樣，系統中的所有主機便可共用連接交換網和端點的結構鏈路。交換網韌體會攔截來自主機的所有配置平面通訊 (包括 PCIe 枚舉過程)，並使用數量可配置的下行埠虛擬化一個符合 PCIe 規範的簡單交換網。

當所有控制平面通訊都路由到交換網韌體進行處理時，資料平面通訊直接路由到端點。其他主機域中未使用的 GPU 不再滯留，因為它們可以根據每個主機的需求動態分

圖 4：雙主機 PCIe 結構引擎



配。結構內支援點對點通訊，這使其能夠適應機器學習應用。當以符合 PCIe 規範的方式向每個主機提供功能時，可以使用標準驅動程式。

操作方法

為瞭解這種方法的工作原理，我們以圖 4 中的系統為例，該系統由兩個主機 (主機 1 採用 Windows 系統，主機 2 採用 Linux 系統)、四個 PAX PCIe 結構交換網、四個 Nvidia M40 GPGPU 和一個支援 SR-IOV 的 Samsung NVMe SSD 組成。在本實驗中，主機運行代表實際機器學習工作負載的通訊，包括 Nvidia 的 CUDA 點對點通訊基準測試實用程式和訓練 cifar10 圖像分類的 TensorFlow 模型。嵌入式交換網韌體處理交換網的低階配置和管理，系統由 Microchip 的 ChipLink 除錯和診斷公用程式來管理。

四個 GPU 最初分配給主機 1，PAX 結構管理器顯示在結構中發現的所有設備，其中 GPU 綁定到 Windows 主機。但是，主機上的結構不再複雜，所有 GPU 就像直接連接到虛擬交換網一樣。隨後，結構管理器將綁定所有設備，Windows 裝置管理員將顯示 GPU。主機將交換網視為下行埠數量可配置的簡單物理 PCIe 交換網。

一旦 CUDA 發現了四個 GPU，點對點頻寬測試就會顯示單向傳輸速率為 12.8 GBps，雙向傳輸速率為 24.9 GBps。這些傳輸直接跨過 PCIe 結構，而無需透過主機。如果運行用於訓練 Cifar10 圖像分類演算法的 TensorFlow 模型並使工作負載分佈在全部四個 GPU 上，則可以將兩個 GPU 釋放回結構池中，將它們與主機解除綁定。這樣可以釋放其餘兩個 GPU 來執行其他工作負載。與 Windows 主機一樣，Linux 主機也將交換網視為簡單的 PCIe 交換網，無需自訂驅動程式，而 CUDA 也可以發現 GPU，並在 Linux 主機上運行 P2P 傳輸。效能類似於使用 Windows 主機實現的效能，如表 1 所示。

表 1：GPU 點對點傳輸頻寬

交換類型	主機 1 平均頻寬	主機 2 平均頻寬
單向 P2P	12.8 GBps	12.7 GBps
雙向 P2P	24.9 GBps	24.6 GBps

下一步是將 SR-IOV 虛擬功能連接到 Windows 主機，PAX 將此類功能以標準實體 NVM 設備的形式提供，以便主機可以使用標準 NVMe 驅動程式。此後，虛擬功能將與 Linux 主機結合，並且新的 NVMe 設備將出現在模組設備清單中。本實驗的結果是，兩個主機現在都可以獨立使用其虛擬功能。

務必注意的是，虛擬 PCIe 交換網和所有動態分配操作都以完全符合 PCIe 規範的方式呈現給主機，以便主機能夠使用標準驅動程式。嵌入式交換網韌體提供了一個簡單的管理介面，這樣便可透過成本低廉的外部處理器來配置和管理 PCIe 結構。設備點對點交換預設情況下處於允許狀態，不需要外部結構管理器進行額外配置或管理。

總結

PCIe 交換網結構是一種能夠充分利用 CPU 巨大效能的絕佳方法，但 PCIe 標準本身存在一些障礙。不過，可以透過使用動態分區和多主機單根 I/O 虛擬化共用技術來解決這些難題，以便可以將 GPU 和 NVMe 資源即時動態分配給多主機系統中的任何主機，進而滿足機器學習工作負載不斷變化的需求。

CTA