# 雲計算網絡虛擬化原理與實現

毛文波，邊瑞鋒，李芳
道裏雲信息技術有限公司，北京
Wenbo.Mao@daolicloud.com

## 摘要

傳統的網絡控制技術基於將一個網絡中的所有網絡包都集中發送至一個叫做 chokepoint 的控制點，在那兒集中處理網絡包元數據（地址、標記、封包、隧道等手段）。由於 chokepoint 是一個物理位置固定的硬件設備，因而用傳統網絡技術構造出的網絡具有很強的位置固定屬性，不能有效滿足雲計算 IT 作爲服務提供方式所需的彈性、分布、按需可變、大規模可擴展、動態遷移、可租戶自定義拓撲等新需求。本文介紹一種適用於雲計算虛擬化數據中心的全新網絡虛擬化技術[1]，可以爲雲計算租戶位於全球任意地點分布式動態租用的虛擬化 IT 資源按需構造出一個雲計算租戶自定義任意拓撲構造滿足任意應用需求的私有邏輯 overlay 網絡。由於該網絡虛擬化技術不再使用集中 chokepoint 網絡控制模型，雲租戶的私有邏輯網絡與數據中心物理 underlay 網絡硬件設備所在的地理位置及其物理屬性徹底無關，不僅租戶邏輯網絡的定義與構造工作可以使用高級語言編程自動化、自助服務方式完成（即，軟件定義網絡 SDN），而且實現了雲租戶私有網絡構造的快速、動態、大規模分布式跨數據中心彈性可擴展、租戶自定義任意拓撲，滿足任意應用需求等有用性質。

**關鍵詞：**網絡虛擬化，軟件定義網絡，SDN，OpenFlow，OpenDaylight，Overlay，Underlay，Network Virtualization Infrastructure NVI，雲計算網絡

## 壹、引言

隨著雲計算的興起，IT 作爲資產擁有的傳統方式正朝著作爲服務租賃的新方式發生轉變，這不僅是我們使用 IT 方式的改變，也是 IT 業務盒子們在“生活方式”上的變化：從原來站在地板、辦公桌等不會思維的支撐物上轉變爲“站在”分布式部署且具有很強智能的虛擬化軟件 hypervisors 集群架構上。讓我們在網絡方面觀察，在這個“生活方式”變化以前，每個 IT 業務盒子都插有一根網線（無線網卡與網線原理相同），網線的另一端是一個專用的網絡控制設備；IT 業務盒子必須通過這個網絡控制設備才有可能與其它 IT 業務盒子發生通信。控制 IT 業務盒子之間通信的手段是在網絡設備上檢查與處理網絡包：允許某些網絡包通過而丟棄另一些網絡包。我們注意到，雖然 IT 業務盒子的生活方式正在轉變爲運行在虛擬化架構軟件上，然而在當今公知的雲計算技術（如 Openstack，2013 年 4 月發布的 Grizzly 版 [2][3]），虛擬化架構軟件仍然還在運用這種傳統的網絡包處理手段進行虛擬機之間通信的控制，即，仍然是將每一個虛擬機的網絡包

發送（當然在新“生活方式”下不可避免通過虛擬網線發送）到一個網絡設備（當然現在這個網絡設備也可以是跑在虛擬化架構上的軟件：虛擬交換機），在那兒集中做網絡包檢查與處理（允許網絡包通過或將其丟棄）。

這種基於網絡包檢查與處理的傳統網絡技術有其從網絡技術歷史發展過程帶來的局限性。首先，在一個集中點檢查與處理虛擬機網絡包這個做法無法利用虛擬化架構分布式處理可以有效提高決策計算性能這一優勢，相反該集中點形成了網絡控制的一個計算瓶頸。其次，基於對網絡數據包檢查的通信控制技術實際上只檢查位於網絡包頭部的一些元數據（地址，標記等），這些元數據通常只占網絡包全部數據的很小一個比例，然而傳統網絡技術的集中控制點由於僅通過從網線中傳來的數據包瞭解 IT 盒子的通信訴求與屬性，於是集中控制點收到網絡包的全部數據可以很大，但也只檢查占比很小的元數據部分，很有可能立即將整個網絡包丟棄，這樣就造成許多無效數據傳輸，卻也是不得已而爲之。這種基於網絡包檢查與處理的傳統網絡控制技術還有多種從歷史技術局限帶來的其它缺陷，且讓我們聚焦討論我們認爲是在雲計算模式下網絡技術中最極待解決的一個。

網絡包頭部地址，標記等網絡包元數據用它們強烈的物理屬性展示了它們在網絡世界中存在的地理位置。IT 業務盒子屬哪一個數據中心，處於該數據中心裏的什麼位置，該數據中心使用了哪一個網絡設備廠商提供的技術或解決方案，等等，都可以明顯地通過 IT 業務盒子通信交互的網絡包中這些元數據得到體現。在 IT 作爲資產擁有的舊時光，網絡所帶有的這些物理屬性既不會給用戶，也不會給設備提供商帶來任何問題或不便。首先作爲資產擁有的 IT，用戶是以靜態方式擁有 IT 資源的，擁有固定不變的網絡地址，不變的網絡拓撲，使用某一家廠商提供的技術不會造成任何使用上的不便。再從設備提供商的角度看，用物理固定方法配置用戶網絡設備當然可以讓設備提供商按用戶的峰值需求銷售設備，這哪會是個問題，何樂而不爲？然而當 IT 轉到了雲上，用戶不僅可以也願意按需動態彈性方式租用 IT 資源，而且還從安全性可靠性角度出發，會希望從不同地域位置分布的雲服務提供商那裏分布式租用 IT 資源。再從雲服務提供商角度出發，也希望 IT 的提供具有可移動、可變化性質，如此不僅能提高數據中心各種硬件資源的利用率，也有利於數據中心對硬件設備的管理與運維。不幸的是，當前雲計算公知技術仍然使用基於網絡包元數據檢查與處理的傳統網絡技術，用這樣的技術形成的租戶網絡不可避免地帶有強烈的物理位置固定屬性，從而很難實現按需、彈性、分布式部署提供的租戶私有網絡，更難滿足 IT 上雲所出現的新需求而必須向租戶提供的可編程、自動化、實時動態滿足用戶變化的租戶私有網絡。雲上租戶的私有網絡應該是一個純粹的邏輯網絡，應該徹底與傳統網絡帶有的強烈物理位置屬性去除耦合，應該將硬件網絡設備徹底池化（pooling）。這種網絡虛擬化的資源池化需求就像當前業已十分成熟的 CPU 虛擬化技術與存儲虛擬化技術已經能夠成功地滿足硬件資源池化情形一樣：租戶只看到自己租用的邏輯 CPUs，邏輯存儲空間所具備的邏輯價值與服務質量（QoS），而絲毫不管不問不顧硬件 CPUs 或硬件磁盤的物理屬性與所在位置。

我們注意到，越來越多的 IT 業務盒子開始運行在虛擬化架構的分布式智能軟件層上這一新“生活方式”可以對 IT 業務盒子的網絡通信控制技術產生一個本質性變化。虛擬化架構對於運行其上的 IT 業務盒子可以施展的控制能力遠大於網絡設備利用網線中傳來的網絡包所能對 IT 業務盒子起到的控制作用。具體地，交換機技術將 IT 業務盒子（節點）的網線集中到網絡設備上（網絡 chokepoint）檢查和處理網絡包元數據這種傳統網絡控制手段可以轉變爲“在網線的虛擬機這一頭”（網線末端 Leaf 節點）實時動態地做網絡控制（比如用“插拔網線”手段實現防火牆控制）。得益於虛擬化軟件架構分布式部署的智能，基於在網線末端實時動態網絡控制處理技術完全可以爲全球分布的任意兩個虛擬機實時動態地處理細化到一根邏輯網線所需的控制顆粒度。請進一步注意到：在虛擬化架構軟件看來，每一個“站立於其上”接受其服務的虛擬機都可以帶有可列無限多個網線插口，每個網線插口上都可以插上一根專用網線，即，虛擬化架構軟件可以爲全球任意兩個虛擬機插上一根專用的邏輯網線，網線中當且僅當傳輸這兩個虛擬機通信交互而產生的網絡包。於是此專用網線的服務狀態完全可以與傳輸於其中的網絡包頭部的 IP、MAC 地址，標記等傳統網絡包元數據所帶有的位置屬性毫無任何相干，比如這種專用邏輯網線的連通服務狀態絲毫不會受到被插上的兩個虛擬機所處的位置動態變化所影響。此邏輯專用網線的定義，存在或消亡可以當且僅當被兩個需要通信的虛擬機之身份屬性，以及它們是否需要通信，在何種條件下，以什麼方式，在什麼時間通信，等這些通信的本原屬性所決定。所以，採用分布式虛擬化架構軟件“在虛擬機這一頭實時動態地處理專用網線中的網絡問題”這種全新的網絡控制技術，我們可以爲雲計算租戶虛擬化出一個純粹的邏輯私有網絡。由於徹底擺脫了網絡的物理屬性，我們的網絡虛擬化工作可以用純軟件高級語言編程方法實現，自動化地、快速地、動態地定義與更新租戶私有網絡，而且如此實現的租戶私有網絡還可以跨數據中心分布式部署。

本文以下部分將詳細介紹道裏雲公司全球首創的網絡虛擬化技術，並用雲租戶分布式虛擬化防火牆應用場景爲例說明該技術的有效性與實用性。

## 貳、技術背景

近年來，隨著網絡互聯與雲計算技術的發展，越來越多的組織、企業或個人正在把 IT 需求從傳統的資產擁有模式逐步轉變爲按需租用服務模式，即，使用“IaaS”（基礎設施爲服務）模式的雲計算。這種雲計算環境由於使用了資源虛擬化技術，不僅租戶可按照之即來，揮之即去方式動態租用資源、可對租用資源作與位置無關的（動態）遷移，而且雲數據中心還可以讓多租戶（multi-tenancy）共享數據中心物理的計算、存儲與網絡資源、可對資源作自動負載均衡，因而大大提高了資源的利用率和服務的可靠性。爲了提高容災可靠性，同時避免 vendor-lock-in （服務商鎖定）問題，一個組織還應該從多個雲服務提供商處租用 IT 資源，這些資源雖然在物理上甚至地理上完全可以分布於不同

的雲數據中心，但從租戶看來就應該像處於一個私有網絡中的內部資源一樣。爲這樣的租戶提供合適的防火牆解決方案是保證租戶 IT 安全（即雲安全）的關鍵。

傳統的（IT 作爲資產擁有的）防火牆產品都基於所謂的 chokepoint 模型：組織先將作爲資產擁有的，物理位置上處於鄰近位置的 IT 資源用網線與硬件網絡設備（如交換機等）互聯，形成一個專屬組織內部私有的本地網絡（LAN），然後設置一個叫做 chokepoint 的集中點，僅在這個點上允許私有 LAN 與公有廣域網（WAN）發生通信，並在這個點部署組織的防火牆內外通信控制策略，形成組織的網絡邊界（network edge）。傳統防火牆的典型拓撲構造如圖 1 所示。在 IT 作爲資產擁有的模式下，組織局域網內部 IT 設備的物理配置（如物理機器的 IP 地址）一般不會發生變化，所以傳統防火牆技術都使用 IP / MAC 地址之類的網絡包元數據作爲參數來定義與配置防火牆允許或禁止通信的控制策略。由於傳統作爲資產擁有的 IT 資源具有物理靜態固定屬性，組織的防火牆及網絡安全策略的配置工作以雇傭專職網絡管理員使用操作系統命令行手段進行手工配置，這種樸素做法完全可以滿足組織作爲資產擁有的靜態 IT 以及相關的靜態網絡安全需求。

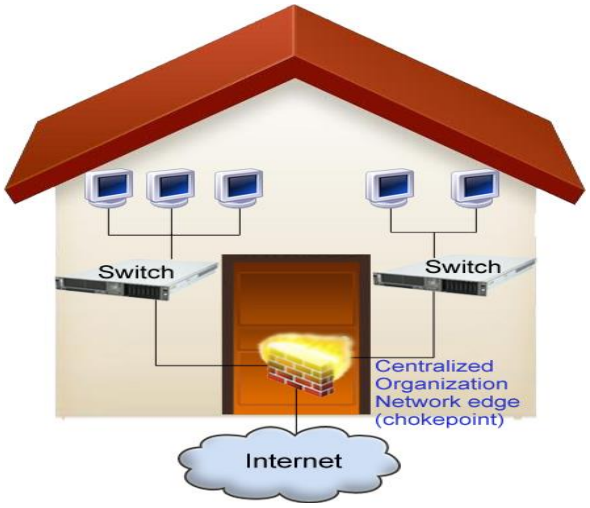


圖 1：傳統組織內部網路與防火牆的 chokepoint 模型

組織的私有網絡與防火牆在雲計算的情形需要轉變爲虛擬化多租賃雲數據中心裏的租戶網絡與防火牆。在當前公知雲計算技術中（如 Openstack 的最新 Grizzly 版[2，3]，於 2013 年 4 月發布，其它如 CloudStack[4]，OpenNebula[5], Eucalyptus[6]）租戶的網絡隨著所租用 CPU 的虛擬化也不可避免地遇到虛擬化問題，然而在當前公知雲計算技術所涉及的網絡虛擬化工作實際上只是做到了交換機和網卡的“虛擬化”，也就是圖 1 中的硬件 switch 盒子到了雲上變成了圖 2 中的“virtual” switch（“虛擬”交換機）或“虛擬”網橋（hypervisor bridge）軟件模塊[7]。其實此類“虛擬”交換機/網橋技術只不過是爲了解決服務器物理網卡的軟件化問題，因爲自從 CPU 被虛擬化後就再也無法用物理手段對作爲軟件的虛擬機插拔網線了，而必須將服務器的物理網卡通過軟件網線聯通到虛擬機上的軟件網卡。這種“虛擬”（軟件是更正確叫法）交換機與網卡技術完全屬 CPU

虛擬化技術的附帶軟件化產物，與 IaaS 所需要的網絡虛擬化技術很不一樣。IaaS 所需的網絡虛擬化技術是將數據中心裏的物理網絡資源“池化”（pooling）爲一個網絡資源池，從而實現可編程網絡，我們將在後面詳細介紹這種需求。

在理解圖 2 工作原理時我們要注意到，與硬件交換機一樣，連接虛擬網卡與服務器物理網卡的軟件交換機所處理的數據仍然是網絡包元數據如 MAC / IP 地址等，即，在處理數據功能上，軟件交換機與物理硬件交換機的功能完全相同。那麼當前公知技術是怎樣爲雲上的租戶提供“虛擬”私有網絡與防火牆的呢？與物理交換機處理網絡包元數據的功能相同，虛擬交換機也可以對指定的網絡包頭部添加新的元數據，這種網絡包元數據添加手段叫做打標記（tagging）。同一租戶所租用 VM 發出的網絡包可以打上同一標記，不同租戶的網絡包打上不同標記，虛擬交換機會按不同標記將網絡包正確地交換傳遞網絡包，因而現實了租戶之間的網絡隔離。圖 2 中所標出的 VLAN 就是一種網絡標記技術。其實 VLAN 之類的網絡包元數據標記技術早已使用在傳統 IT 作爲資產擁有的情況，比如圖 1 所示的兩個子網（隔離同一組織中的不同部門）大多都是用 VLAN 技術分割而得的。

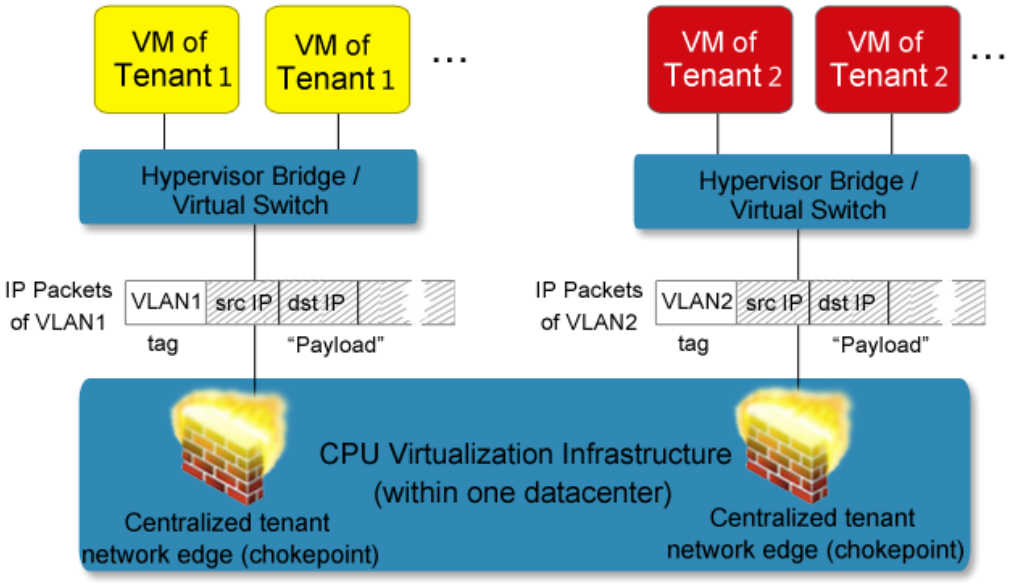


圖 2：多租賃雲資料中心裡租戶網路與防火牆，仍然屬於 chokepoint 模型

爲滿足雲計算網絡可編程的新需求，業界還提出了 OpenFlow 交換機與協議技術[8]，由 Open Network Foundation ONF 作爲未來網絡新標準嘗試商用推廣，如圖 3 所示。OpenFlow 交換機與協議技術具有如下背景：網絡可編程的需求由來已久，傳統網絡設備廠商都採用專用硬件設備內置封閉產權技術的非開放手段使網絡交換機在一定程度上可編程，更準確的說法是讓網絡硬件設備可被編程方法配置。網絡設備配置技術的非開放性的一個明顯特徵就是網絡控制邏輯（位於控制平面）與數據包轉發邏輯（位於數據平面）都是零散部署在各個網絡硬件設備中的。顯然這種零散部署在各個網絡硬件設備中

的網絡配置方法根本無法滿足雲計算大規模動態快速網絡變化所需的數據中心全域分布的網絡甚至跨數據中心分布網絡規模的可編程需求。OpenFlow 交換機與協議技術提出全新的下發指令可編程交換機原理（與存儲指令可編程計算機原理相同），從一個邏輯上集中的控制器通過協議技術向 OpenFlow 交換機、路由器下發含有多元組格式的網絡控制指令，OpenFlow 交換機、路由器收到多元組指令後根據指令對給定的網絡包元數據頭部做增、刪、查、改等操作，這樣的操作構成了對交換機、路由器的編程，從而使交換機、路由器按照指令發送、接收或處理網絡數據包。比如，一條下發至某個交換機的多元組數據可以指令該交換機爲某些網絡包打上給定的 VLAN 標記，用以實現網絡隔離。OpenFlow 以這樣的方法實現了軟件定義的網絡 SDN。

我們可以看到 OpenFlow 技術企圖對現有網絡世界裏的物理硬件基礎設施來一個改朝換代式的升級更新。從商業投資與回報的關係來看，假定交換機升級可以像“革命”那樣一夜發生，也只可能最先發生在那些已經在利用網絡流量直接產生商業回報的服務提供商所用的網絡基礎設施上。可以合理地估計此類服務提供商多爲電信運營商，因此最早在商業上得到應用的 OpenFlow 網絡設備應該是運營商的底層核心級交換機與路由器，而非雲計算數據中心裏大量採用的接入級交換機設備。我們還可以進一步估計：電信運營商所關心的 SDN 問題更多集中在如何對主幹網實施流量工程（Traffic Engineering，TE）方面，即，發現和利用主幹網上最大流量的路徑，而雲計算服務提供商則更關心如何對個體租戶提供面向應用的網絡實現 QoS（Quality of Service）增值，比如解決租戶隔離，防火牆，入侵檢測，DDoS 規避與流量清洗，網絡負載均衡，等問題。

OpenFlow 標準技術當然也可以使用基於開放源代碼實現的虛擬交換機，虛擬交換機可以使用 OpenFlow 協議實現各種控制功能，OpenvSwitch 開源項目[9]允許第三方網絡技術提供商，如 Nicira[10]，NEC[11]，Ryu[12]，的技術以 Plug-in 模塊方式使用 OpenFlow 交換協議控制虛擬 OpenFlow 交換機。然而我們必須注意到 OpenFlow 格式的網絡包不能獨立僅在虛擬交換機組成網絡中流動，因此可以把 OpenvSwitch 項目的最終目的看作是對 OpenFlow 硬件網絡設備升級換代完成後的補充。

暫且不論 OpenFlow 通過對網絡基礎設施做改朝換代式的升級路徑能否快速在雲數據中心網絡上得到商業應用，我們注意到，OpenFlow 對網絡設備 SDN 編程最終達到的目的仍然是通過處理網絡包元數據的傳統手段，即，地址、標記、封裝、隧道，實現網絡控制。這與傳統設備廠商封閉技術的做法殊途同歸。既然“革命”的結果仍然離不開傳統網絡包元數據處理，漸進式的革新或許比改朝換代式的“革命”更容易成功，於是以傳統網絡設備廠商爲主，業界又於 2013 年成立了 OpenDaylight 開源項目[13]，不是通過設定“革命”新標準對網絡基礎設施做改朝換代，而是通過開放源代碼途徑僅解決以前各家封閉做法所具有的可擴展性差的問題。

網絡的大規模動態可擴展性應該是當前雲計算網絡極待解決的一個重大問題，只有可以並網擴展的雲網絡才可能讓雲計算具有彈性按需服務性質，這與不可並網的發電技術不具有商用市電服務價值是一個道理。雲計算數據中心必須以規模化多租賃方式運

營。如果僅採用傳統 VLAN 技術隔離租戶，則由於 VLAN 標準採用 12 個比特標記 VLAN 身份，一個數據中心至多容納得下 4096（=2^12）個可以購買安全隔離作爲服務的租戶。於是 VMware 採用其收購的 Nicira 網絡虛擬化技術於 2013 年 8 月推出了 NSX 網絡虛擬化技術[14]，採用 VXLAN 標準可以將 VLAN 身份標記擴展至 16M（約 1 千 6 百萬=2^24）規模。考慮到雲計算租戶防止服務商鎖定（Vendor Lock-in）風險需求，VXLAN 標準需要在跨雲服務提供商範圍應用，全球範圍租戶數量，尤其考慮到長尾小租戶，將遠遠超過 16M，所以我們認爲 VXLAN 標準不足以作爲未來雲計算網絡可擴展標準。

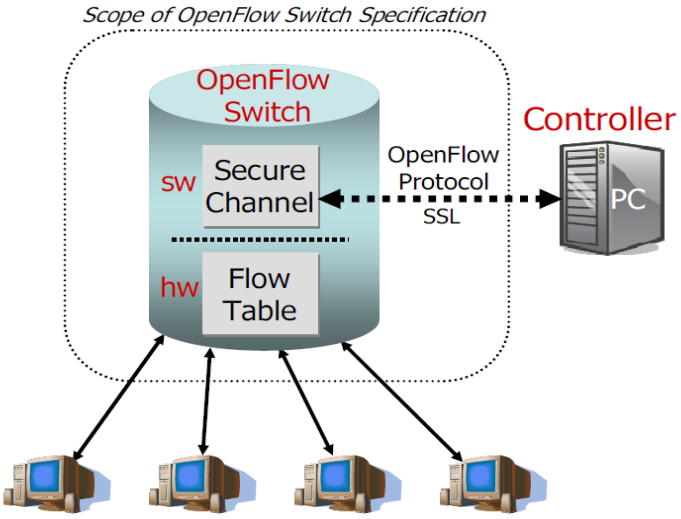


圖 3 ：OpenFlow 技術，仍然屬於 chokepoint 集中網路包中繼資料處理模型

這裏所綜述分析的網絡技術皆採用處理網絡包元數據，即，地址、標記、封裝、隧道等手段實現網絡控制。歷史上 IT 業務盒子都必須通過網線（包含無線形式的網線）將網絡包送到交換機才能在那兒處理網絡包元數據，比如 OpenFlow 新標準嘗試情況，圖 3 中連接終端節點與 OpenFlow 交換機的連線都是網線。交換機通電後的首要任務就是通過網線學習由 IT 業務盒子的網卡組成的網絡，每一個交換機都構成一個其學習所得網絡通向外部世界的唯一控制點，叫做 chokepoint，發自該網絡內部的網絡包一律先上行發送到該 chokepoint 根據通信策略對網絡包實施處理，同理，發自該網絡外部的網絡包也一律先發送到該 chokepoint 根據通信策略對網絡包實施處理。比如網絡防火牆情況就是在該 chokepoint 做如下包處理：ACCEPT（通過），即允許外發或下行內發，或 DROP（丟棄），即禁止外發或下行內發。我們可以將本章所綜述的 IT 網絡控制技術模型概括描述爲：在網絡 chokepoint 處集中處理網絡包元數據模型，該模型可以簡稱爲“網絡 chokepoint 集中控制模型”。

由於網絡 chokepoint 集中控制模型帶有交換機帶有的物理（比如位置）屬性，基於該模型的網絡不容易被虛擬化，也不容易按動態分布式大規模可擴展方式部署。比如觀察 chokepoint 集中網絡包元數據處理的典型用例：在 VLAN 隧道標準上應用 VXLAN 封包標準擴展虛擬化網絡規模，所得規模僅限於容納 16M 個租戶。又觀察在該模型下實現

的防火牆：通常網絡包元數據僅占有整個網絡包全部數據（包頭元數據 + Payload 上層數據）很小比例，但該模型下的防火牆必須先將整個網絡包全部數據發送至網絡 chokepoint，在那兒通過檢查元數據判定丟棄掉或轉發，因而在丟棄情況會產生大量無效網絡流量。

## 參、雲計算網絡虛擬化需求

對網絡技術的發展途徑討論至此我們知道，雖然當前雲計算採用的公知網絡技術已經將專用的，封閉產權的，用物理硬件提供的交換機技術開放轉變成爲跑在普通大宗服務器上的，用開放源代碼實現的軟件交換機技術。但是該軟件所處理的數據仍然是網絡包元數據，仍然像 IT 作爲資產的傳統網絡技術情況那樣：租戶之間的網絡隔離是通過對網絡包打上不同租戶標記，因而可做區分處理，實現的。不僅如此，租戶防火牆的控制也仍然與 IT 作爲資產情況一樣：先將租戶網絡包一律發送至一個集中的 chokepoint 控制點做網絡包元數據（地址、標記）檢查，根據租戶設定的通信策略決定是否允許轉發，以這樣的方法實現控制內、外通信。這些做法都與圖 1 的傳統組織私有網絡與防火牆的工作原理完全一樣，仍然屬 chokepoint 模型。這種傳統的網絡與防火牆技術無法解決 IT 轉到雲上 IaaS 遇到的如下新問題：

1) **租戶新需求：**如何向租戶提供可由租戶自定義網絡拓普構造的，安全隔離的、動態彈性的、分布式部署的、租戶私有網絡。雲計算租戶希望用召之即來，揮之即去方式租用雲計算資源，因此具有：網絡動態彈性可擴展、高可靠、高可用、高安全、強隔離、免除 vendor-lock-in 風險，等新的合理需求。這些租戶新需求在技術上的體現，或者說在學術上有意義的描述，就是網絡拓撲與控制的安全分布式計算：需要針對跨多個數據中心，跨多個服務提供商構造邏輯上私有的租戶網絡。這樣的網絡及安全策略的配置和實施必然是一種分布式的網絡與防火牆架構，傳統防火牆的集中式 chokepoint 模型顯然無法滿足這些新需求。以 VLAN 之類的網絡包標記技術爲例，這種傳統網絡標記技術與物理網絡底層構造緊密耦合，不僅不能跨數據中心部署，而且不同的網絡設備廠商都有自己特殊的接口標準，十分突出地具備“vendor-lock-in”缺陷。

2) **雲計算數據中心新需求：**如何滿足雲計算數據中心的剛性需求，如動態彈性地對雲數據中心資源（包括計算，網絡，存儲）實現有效利用，比如爲負載均衡，容災備份而將虛擬機鏡像文件及租戶數據實施網絡/異地存儲。考慮雲數據中心自動負載均衡需求，租戶虛擬機默認是可動態遷移的，於是如果用網絡包元數據處理技術來隔離租戶虛擬網絡，則不可避免要解決以下兩種情況之一：（i）或者

允許虛擬機的 IP 地址等網絡包元數據隨虛擬機位置的變更而變更，因而相應地動態變更網絡 chokepoint 控制點的安全配置參數，（ii）或者用某種技術鎖定虛擬機的網絡包元數據，使之與虛擬機位置的變化無關。還是以 VLAN 方法爲例，此類網絡包元數據標記技術屬（ii）中處理方法，與物理網絡底層的硬件屬性緊密耦合，因而在租戶網絡的可擴展性、動態彈性、尤其在可分布部署方面，都具有很差的性能。又以 chokepoint 模型的防火牆邊界控制爲例，這種典型的集中式計算模型天然不具備分布式部署的可能。

雲計算 IaaS 所需要的網絡虛擬化（也是本產品說明書將要介紹的網絡虛擬化技術）應該是將數據中心裏的物理網絡資源虛擬化，也叫做網絡資源池化（network resource pooling），這種資源池化的結果應該將租戶網絡變成一個純粹的邏輯網絡，徹底與數據中心物理網絡資源的硬件物理屬性無關（這種無關性又叫做去除耦合，de-couple）。在一個徹底與物理網絡去除耦合的邏輯網絡上做 QoS（分布式租戶防火牆就是一種較高級的網絡 QoS）增值工作可以用純軟件編程技術完成，實現自動、快速、動態的解決方案，編程工作無需考慮網絡的物理屬性。比如在部署分布式租戶防火牆情況，無需考慮在哪個數據中心部署租戶的防火牆 chokepoint。又比如，爲了判斷兩個節點是否能相互通信，一個去除了耦合的邏輯判斷方法不應該將大量數據包發送到物理網絡的一個集中點，那樣做不僅會造成不必要的無效網絡流量，還會在該點造成集中式的計算處理負擔。

因此，實現雲計算平臺虛擬網絡與數據中心物理網絡的徹底去耦合，實現網絡硬件設備資源的徹底池化，實現租戶私有網絡的徹底虛擬化，構建跨越數據中心的網絡虛擬化基礎設施（Network Virtualization Infrastructure，NVI）是雲計算 IaaS 架構的一個關鍵技術。雲計算網絡虛擬化以及雲安全所需的租戶防火牆當前技術所處的較爲初級現狀要求我們尋求新的，更爲有效的技術手段。

除此之外，現有的“IaaS”雲計算技術還可能面臨如下安全問題[15—23]：

1) 如何保證部署在雲計算平臺上的租戶組織內部信息的傳輸與存儲的安全性？包括防止已經成功入侵的駭客或惡意代碼盜走組織內部數據，防止雲計算平臺服務提供方內部（如系統運維管理）人員有意或無意拷走數據（系統管理員由於具有很高的權限，可以通過諸如網絡抓包，拷貝虛擬機鏡像文件等手段，有意盜取重要數據）。
2) 如何防止組織內部用戶惡意洩漏組織內部數據？比如向互聯網發送數據，或通過終端上的移動存儲介質口向外拷貝文件等等。
3) 如何在雲平臺網絡環境內建立可靠、可信的計算機數據管理安全策略與機制？如提供基於策略的數據流向控制，要求數據只能從外網單向流入內網，或只能從低級別節點流向高級別節點。

4) 如何實現權限分立的系統結構？比如隔離不同租戶，或在同一租戶內部隔離不同業務部門，並提供不同安全級別的服務隔離或單向性數據流動機制。

道裏雲公司自主創新，全球首創研發的網絡虛擬化技術設施 NVI 技術通過解決雲計算環境所需的租戶邏輯網絡軟件可編程需求，用一種顛覆性創新手段解決雲計算環境中邏輯網絡與物理網絡徹底去耦合難題，使物理網絡資源徹底池化，使租戶網絡的各種 QoS 增值編程任務所需的解決手段與底層物理網絡構造完全徹底無關，從而使租戶網絡中任何 QoS 問題可以完全徹底轉變爲高級語言軟件編程問題自動快速解決。得益於此種網絡去耦合問題的解決，本產品完全用高級語言軟件編程方法，實現了租戶跨數據中心分布的，自助服務方式自定義的，具有任意拓普構造的，租戶虛擬化網絡與防火牆。

## 肆、道裏雲網絡虛擬化架構技術原理

在服務器虛擬化技術以前的 IT 業務盒子都站在地上、桌上、或其它種種不能思維的介質上，若要對 IT 盒子之間的通信做控制須先將它們一一連接到一些中介於它們之間的網絡設備。除特殊情況之外，我們可以說在全球任意兩個 IT 業務盒子之間都必然中介有至少一個網絡設備。在這些中介於 IT 盒子之間的網絡設備上可以部署通信控制策略，或允許發送端的網絡包轉發至接收端，或直接將網絡包丟棄。傳統的 IT 通信技術就是通過在中介網絡設備上如此處理網絡包方法，可以控制位於全球任意位置的兩個 IT 業務盒子之間的通信。在不失一般性的情況下發送端與接收端分別屬兩個不同的網絡（即，由兩個交換機各自學習所得網絡），網絡包從發送端至接收端須通過網絡 chokepoint，因此傳統網絡控制邏輯都部署在網絡 chokepoint 設備上。

雲計算服務器虛擬化技術使 IT 業務盒子的運行環境發生了根本性變化：IT 業務盒子越來越多地變成了虛擬機，“站在”分布式軟件 VMMs（hypervisors）集群上。與以前 IT 業務盒子下面的地板或桌子等非智能支撐物不同，虛擬機“所站立”的 VMMs 具備很強的智能，完全可以利用他們來控制虛擬機之間的通信。進一步觀察，以前中介於 IT 業務盒子的網絡設備是通過網線連接 IT 業務盒子的，通過網線，網絡設備頂多就只能得到 IT 盒子發送的網絡包而已，所以 chokepoint 模型控制通信手段也就只能通過處理網絡包元數據方法進行。然而在虛擬化架構情況，VMM 對於跑在其上的虛擬機所知曉的屬性，以及可以對虛擬機實施的控制，遠多於並強於在網絡集中 chokepoint 點處區區處理網絡包元數據的手段：比如 VMMs 可以知曉虛擬機之身份屬性，還可以對虛擬機就地“插拔”虛擬網線。於是我們不難做出論斷：在虛擬化架構上如果還按以前地板上的生活方式，先將 IT 業務盒子的網絡數據包上行發送至某個集中網絡設備 chokepoint 控制點，在那兒決定是否允許外發或下行，這種典型的慣性思維不僅無法突破傳統通信控制技術由於與集中 chokepoint 點的物理硬件設備緊密耦合所而具有的位置局限性，集中計

算模型的非動態可擴展等弱點，更糟糕的是它完全忽略了 VMMs 所具備的分布式的、超強的智能性。

道裏雲網絡虛擬化架構（DaoliCloud Network Virtualization Infrastructure, NVI）就是基於上述重要觀察，利用分布式 VMMs 集群超強的智能性與靈活的控制能力實現對虛擬機之間通信的控制。NVI 架構爲每一個虛擬機定義了一個全球唯一可標識、可識別的身份 ID。雲計算租戶可以在全球範圍的數據中心分布式租用虛擬機，並且用這些虛擬機的身份自助定義私有的分布式網絡拓撲、組織內部邏輯通信策略、以及分布式防火牆邊界控制策略。這些租戶自助定義的網絡拓撲及通信策略都列出於一個同樣可以在全球範圍分布式部署的數據庫系統中。NVI 位於全球分布的 VMMs 集群可以與數據庫系統交互，按照租戶給出的網絡定義爲租戶的任意一對虛擬機做實時動態地“插拔網線”，從而實時動態地爲租戶全球分布的虛擬機構造出一個租戶私有網絡。圖 4 形象化地描述了網絡虛擬化 NVI 架構的技術原理。

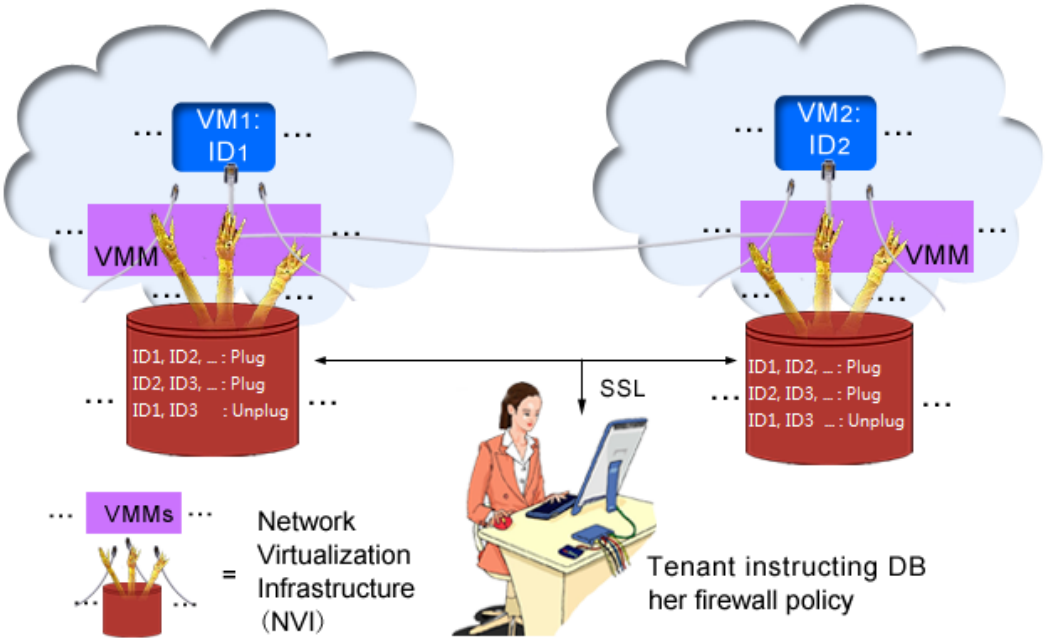


圖 4： NVI 原理：為任意兩個虛擬機“插拔專用網線”的 Leaf 網络控制模型

我們可以把 NVI 網絡虛擬化架構想成一個智能的“千手觀音”，它聽從數據庫指令（數據庫中記錄了租戶指令），可以對全球任意地點分布的兩個虛擬機按照租戶指令實時動態地“插拔網線”，而且對“插上的網線”還可以做應用所需的 QoS 增值處理。用這樣直接而簡單的原理，NVI 可以構造出一個租戶自助定義的，可以全球分布式部署的虛擬私有網絡。我們格外要注意的是：NVI 架構爲一對虛擬機所插上的網線是這兩個虛擬機專用的，該專用（邏輯）網線中傳輸的所有網絡包都當且僅當只屬這兩個虛擬機，沒有任何其它第三方的網絡包會在此專用網線上通過。由於全球分布的數據中心服務器都具有固定的物理地址（如 IP 地址），所以我們可以認爲全球分布的服務器在物理上都是連接

著的，比如可以應用 VPN 技術將它們聯通。於是我們可以讓物理上已經相連的一對 VMMs 根據數據庫中的指令，用軟件方法爲兩個虛擬機“插上”一根專用網線，並加以應用所需的 QoS 處理。從圖 4 我們可以看到，用 NVI 技術實現的網絡控制不再屬傳統的網絡 chokepoint 集中式控制模型，網絡控制發生在 IT 業務盒子端，所以 NVI 技術屬網絡末端（Leaf）分布式控制模型。由於末端控制針對一根邏輯專用網線進行，僅需該專用網線所連接虛擬機的身份，無需涉及任何網絡包地址信息。正是基於這個原理，NVI 網絡虛擬化架構完全避免了對網絡包元數據做任何處理或關心，而僅用 IT 業務盒子的身份屬性決定它們是否能通信。所以由 NVI 實現的邏輯網絡與底層支撐的物理網絡所帶有的硬件設備屬性、以及位置屬性等徹底去除了耦合。我們認爲 NVI 架構是對網絡虛擬化技術的一個新發展，其意義就在於，由於 NVI 構造的租戶邏輯網絡避免了對網絡包元數據的處理，完全屏蔽掉了網絡包元數據（MAC、IP 地址，網絡包頭部標記等數據）多變的屬性（網絡包元數據的多變屬性與租戶動態變化，數據中心負載平衡機制等需求相關），於是租戶私有網絡的 QoS 增值工作可以容易地由純軟件方法定義（SDN），可以實現自動化方式動態快速部署。

圖 4 所描述的“爲任意兩個虛擬機插拔專用網線”，其本質內容爲：將虛擬機邏輯身份實時動態地映射（map）到虛擬化架構底層物理網絡資源。這個映射從一個虛擬機生成時刻開始，在虛擬機整個生命周期中持續得到維護，以下是描述該映射的 NVI 算法，該算法使用 IPv6[24]地址作爲虛擬機永遠不變的邏輯身份，將這個邏輯身份動態映射至一個經常變化物理 IP 地址，物理 IP 地址經常變化的原因是虛擬機運行位置發生了動態遷移。

**NVI Algorithm: Network Virtualization for the Whole Lifecycle of a Virtual Machine**

INPUT: A tenant T who has initialized the tenant entry in the DB of a VI;

OUTPUT: A VM over the VI, which is rented by T with a unique IPv6 address registered in the DB of the VI, where the IPv6 address is always mapped to an ephemeral physical network address

1) Upon T request to create a VM, NVI chooses a resource available hypervisor H; let H create VM with a globally uniquely identifiable id, let V6 denote this id;
2) Let H create in DB a new entry=V6 for the newly created VM; this V6 entry is added to the tenant entry;
3) Let H record in DB by adding the current network address of VM over H to the V6 entry; (let Physical-IP denote the current network physical address of the VM over H)
4) (The cryptographic functions in following steps are optional for providing cryptographic protection to the VM ID) Applying public-key cryptography, let H create a PKI certificate Cert(V6) and a digital signature Sign(V6, Physical-IP) for VM

in such a manner that the correctness of the mapping (V6, Physical-IP) can be cryptographically verified by any entity using Cert(V6) and Sign(V6, Physical-IP);

5) Let H record in DB by adding Cert(V6), Sign(V6, Physical-IP) to the V6 entry;
6) Upon motion of VM (live or static migration), let a destination hypervisor DH in NVI take over the DB maintenance job for VM; let Physical-IP' denote the new network address for VM over DH; let DH update Sign(V6, Physical-IP') in the V6 entry to replace Sign(V6, Physical-IP);

**End of NVI Algorithm.**

NVI網絡架構的末端控制模型可以爲一個給定IT業務盒子集合做出一個具有任意拓撲，任意性質的overlay邏輯網絡，與網絡底層物理硬件設備資源的屬性完全脫離關係，因此網絡底層物理硬件設備資源被徹底地池化了（network resources pooling）。這種邏輯網絡與物理網絡去除耦合的關係類似於 CPU 虛擬化架構技術實現了對物理硬件 CPU 徹底資源池化情形：IT 通信究竟使用了位於何處的哪一根網線和哪一個廠商的設備，這些都無關緊要，雲租戶只關心自己定義的虛擬私有網絡在邏輯上具有所需的 QoS 性能。

## 伍、道裏雲網絡虛擬化架構 NVI 算法實例：虛擬化分布式防火牆

以上我們介紹了網絡虛擬化架構 NVI 的 leaf 節點網絡控制技術原理，現在讓我們就租戶邏輯網絡 QoS 的增值服務爲租戶虛擬化分布式防火牆這一應用場景情況，具體察看 NVI 如何爲一個雲租戶實現分布式虛擬私有網絡與防火牆。這種分布式私有網絡與防火牆的典型網絡分布場景如圖 5 所示：

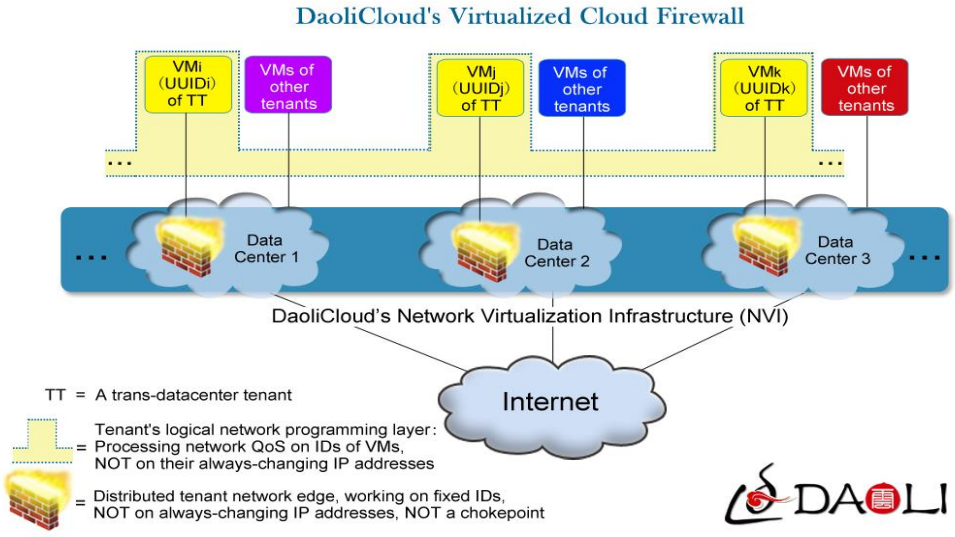


圖 5：租戶虛擬化私有網絡與防火牆：網絡的分散式運算

當租戶通過自助服務方式爲全部租用的虛擬機的身份定義完畢通信策略後，這些基於虛擬機身份定義的通信策略，以及用於證明虛擬機身份真實性的 PKI 證書，都被 NVI 架構的數據庫管理系統所管理與維護。圖 6 示例了道裏雲 NVI 架構操作系統的虛擬機管理頁面，該管理頁面顯示每當系統新生成一個虛擬機時，會爲虛擬機同時生成一個全球範圍唯一標識的身份，以及爲證明該身份的真實性而生成的一個 PKI 公鑰證書：

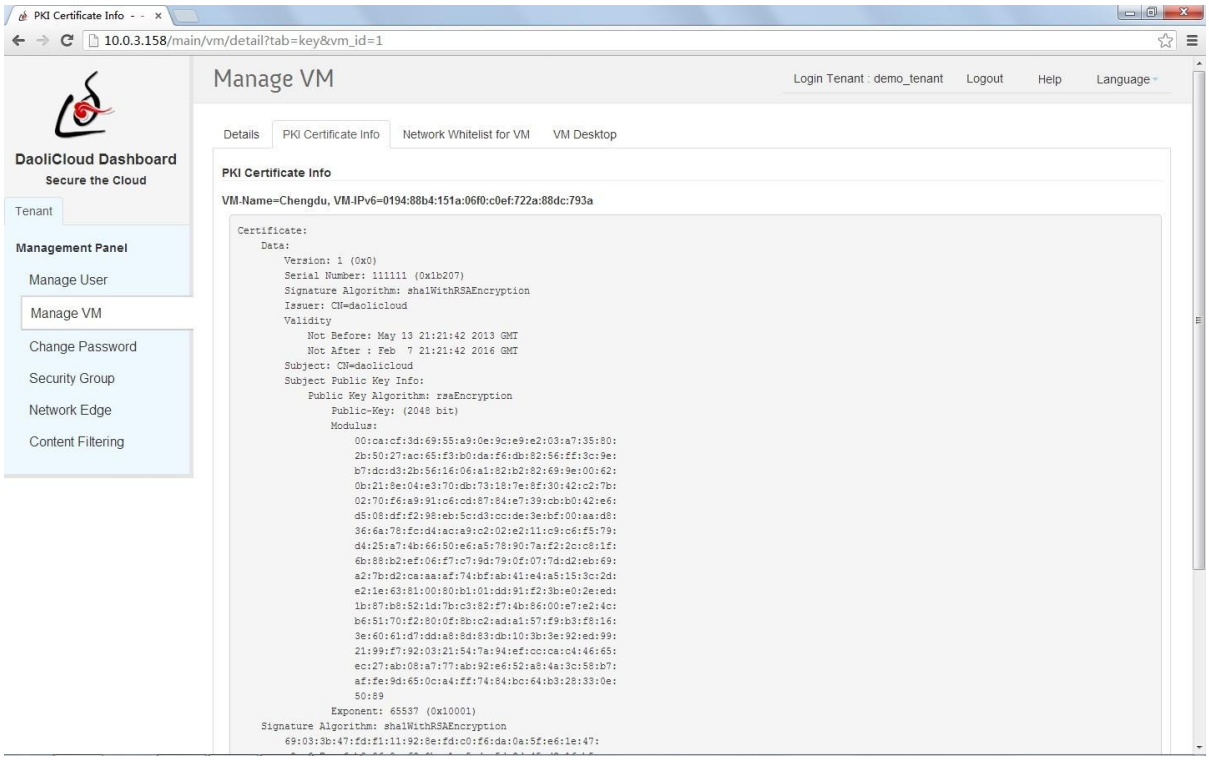


圖 6：虛擬機唯一 ID 身份及 PKI 證書

以下我們介紹當租戶的兩個虛擬機發生通信時，租戶虛擬網絡防火牆的 SDN 算法。

**SDN Algorithm 1 (Sender's Protocol): Cloud Tenant's Network and Firewall Definition**

INPUT: Let VM1 of ipv6-address=SRC on hypervisor SH be a communication source to establish a communication session with VM2 of ipv6-address=DST on hypervisor DH; Let DB contain a tenant's firewall policy P;

OUTPUT: VM1 is inside/outside the tenant's firewall according to the policy P;

1) Let SH intercept a network packet of source from VM1; let SH use the destination ip address DST in the packet to search in DB for the mapping physical address of the destination;
2) Let DB use DST to find Cert(DST), Sign(DST, Physical-DIP); let DB also use SRC to find and the tenant's communication policy for SRC; let DB return the search result to SH;

3) Let SH verify the correctness of Cert(DST), Sign(DST, Physical-DIP); if the verification returns NO, SH terminates service (unplug at the sender's end);
4) (The signature verification in Step 3 returned YES) Let SH verify the tenant's communication policy P for SRC; if P contains "SRC, DST: Unplug", then SH terminates service (unplug at the sender's end);
5) (P contains "SRC, DST: Plug") Let SH plug the unicast cable for SRC;
6) (The cryptographic description in the following steps are optional for providing cryptographic protection on the network traffic) Let SH initiate a cryptographic protocol (e.g., IPsec) with DH to provide a cryptographic protection on application layer data in the network packet;

**End of SDN Algorithm 1.**

我們要注意到，當通信的兩端都跑在 NVI 架構上時，以上 SDN 算法的輸入數據中不含有任何網絡包源數據（地址或標記）信息，具體地，插拔網線算法步驟（1）中，源 SH 查詢 DB 時所用的目的地址 DST 是目的虛擬機 VM2 永遠不變的邏輯身份，而步驟（2）DB 返回的 Physical-DIP 則是目的虛擬機 VM2 當前所在位置的臨時物理地址，該臨時物理地址既不作爲算法的輸入，也不是算法處理的數據，而且也根本不出現在租戶定義的通信策略中（租戶自定義的通信策略定義在虛擬機永久不變的邏輯身份上）。在具體實現上，該臨時物理地址是由服務於目的虛擬機 VM2 的目的 VMM 基於上一章介紹的 NVI 映射算法臨時動態分配給 VM2 的。已知網絡技術（如 DHCP 算法）可以爲虛擬機分配與動態管理臨時的網絡地址。讀者僅需明白，此臨時物理地址的變化與否不會對算法的結果造成任何影響。所以用 SDN 防火牆算法構造而得的租戶組織內部私有網絡是一個純粹定義在虛擬機身份上的網絡，與虛擬機所處的物理位置，網絡地址，所用的網絡設備，等物理網絡屬性絲毫沒有任何關係。這就是 NVI 架構上網絡虛擬化的純粹邏輯屬性，與物理網絡去除耦合的徹底性。

其次我們還要注意到，使用以上 NVI 插拔網線算法構造的租戶私有邏輯網絡中不存在"網絡包上行發送"至一個集中 chokepoint 控制點，該邏輯網絡也不存在一個集中的 chokepoint 控制點：首先對租戶內部虛擬機"插拔網線"是在相關虛擬機的當前服務 VMMs 中分布式 leaf 節點處就地處理的；其次租戶的網絡邊界也是在租戶虛擬機端就地"插拔網線"而實現的，所以租戶網絡邊界也是分布式的。這就是 NVI 架構上網絡虛擬化的可分布部署之有用性質。傳統網絡控制技術中"將網絡包上行發送至一個集中點做控制處理"的做法不再適用，分布式"就地插拔網線"的通信控制方法不僅大大減少了無效網絡數據傳輸，提高了傳輸效率，而且還降低了集中式處理天然帶有的單點失效風險。

SDN 算法 1 定義了源 VMM 的通信協議執行步驟，以下 SDN 算法 2 是對應的目標 VMM 通信協議防火牆 SDN 算法：

**SDN Algorithm 2 (Recever's Protocol): Cloud Tenant's Network and Firewall Definition**

INPUT: Let VM2 of ipv6-address=DST on hypervisor DH be a communication destination to respond to a communication session initiated by VM1 of ipv6-address=SRC on hypervisor SH; Let DB contain a tenant's firewall policy P;

OUTPUT: VM2 is inside/outside the tenant's firewall according to the policy P;

1) Let DH intercept the network packet of destination to VM2; let DH use the source ip address SRC to search in DB for the mapping physical address of the source;
2) Let DB find Cert(SRC), Sign(SRC, Physical-SIP), and the tenant's communication policy; let DB return the search result to DH;
3) Let DH verify the correctness of Cert(SRC), Sign(SRC, Physical-SIP); if the verification returns NO, DH terminates service (unplug at the receiver's end);
4) (The signature verification in Step 3 returned YES) Let DH verify the tenant's communication policy P; if P contains "SRC, DST: Unplug", then DH terminates service (unplug at the receiver's end);
5) (P contains "SRC, DST: Plug") Let DH plug the unicast cable for DST;
6) (The cryptographic description in the following steps are optional for providing cryptographic protection on the network traffic) Let DH respond the cryptographic protocol (e.g., IPsec) to DH to provide a cryptographic protection on application layer data in the network packet;

**End of SDN Algorithm 2.**

有興趣的讀者讀者還可以考慮當通信的任意一端是傳統的非虛擬化 IT 盒子的情形，那當然是租戶爲其所租用的虛擬機定義組織網絡邊界的情況。NVI 架構要求 IT 業務盒子中至少有一端是運行在 NVI 架構上的，從而可以在虛擬化架構上實時動態地爲虛擬化的 IT 業務盒子"插拔網線"。

SDN 防火牆算法中步驟（6）中的 IPsec[25]處理選項是對租戶網絡進一步實施基於密碼學力度保護：在"插網線處"調用 IPsec 網絡安全協議標準，該密碼學處理的原理如圖 7 所示。

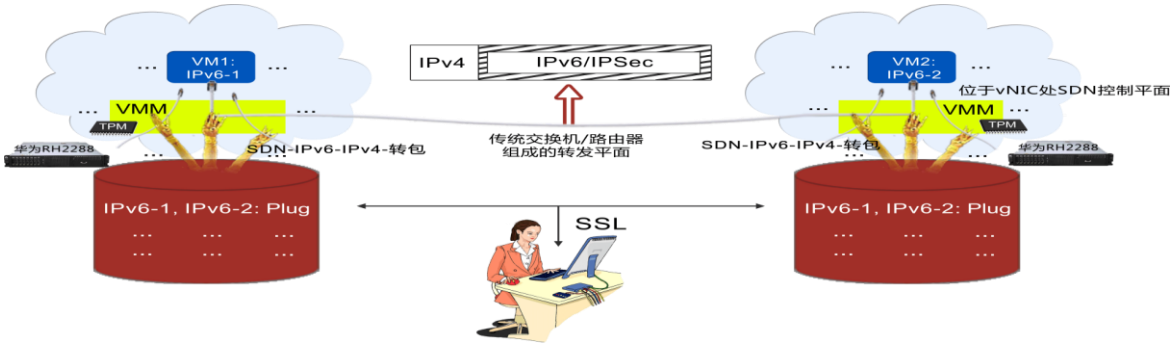


圖 7：IPsec 協議的 SDN 處理

圖 7 所示的 SDN 算法解决了業界兩個知名的網絡難題：

1) 在虛擬網卡處（“插拔虛擬網線處”）用 SDN 編程技術將虛擬機相互通信的 IPv6 格式包封裝入 IPv4 格式包（收包虛擬網卡處做逆運算，拆開 V4 包得到 V6 包），利用現有的 IPv4 交換、轉發平面物理網絡基礎設施跑通 IPv6 標準，可以促進 IPv6 標準早日實現大規模應用。將 V6 包封裝入 V4 包在技術上是個簡單問題，既可以在客體操作系統中做，也可以在 chokepoint 網絡設備中做。這兩種做法都屬對整個基礎設施做更新換代的問題：前者須對全球操作系統做升級，後者須對全球網絡硬件設備做升級換代。此類升級換代的過程需要若干年代才能完成（IPv6 標準於 1998 年提出，具 Google 統計，全球使用 V6 的用戶數於 2003 年 9 月達到了 2%[24]）。道裏雲 NVI 技術在虛擬化架構上的 SDN 做法不僅無需等待基礎設施更新，相反可以促進 IPv6 標準早日實現大規模部署應用。
2) 被 IPsec 協議密碼學技術保護的不同租戶網絡包可以安全共享底層傳統物理網絡資源，比如讓諸多長尾小租戶安全地共享一個 VLAN，不同租戶的網絡依然享受基於密碼學強度保護的安全隔離。如此可以大量節省底層物理網絡資源，雲計算租戶個數不會受到基於現有物理網絡標準實現的網絡規模限制。作爲比較，VMware（Nicira）最近推出的 NSX 網絡虛擬化技術[14]在網絡大規模可擴展方面採用了物理網絡並網標準 VXLAN 技術，因而可以享受安全隔離的雲計算租戶個數不得超過 1 千 6 百萬（16M）。進一步，如果採用 TCG 可信計算技術[26]保護與管理 IPsec 密鑰，則基於密碼學保護的租戶安全隔離還可以防止數據中心系統管理員或成功入侵到系統管理平面的攻擊者通過簡單抓包手段破壞隔離。

當雲計算普及之日，大多 IT 業務盒子變成虛擬機之時，爲全球任意兩個 IT 業務盒子動態實時“插上”一根專用網線將成爲輕而易舉的現實。細顆粒顆網絡流量的按需使用的付費計算也將變得十分容易實現。

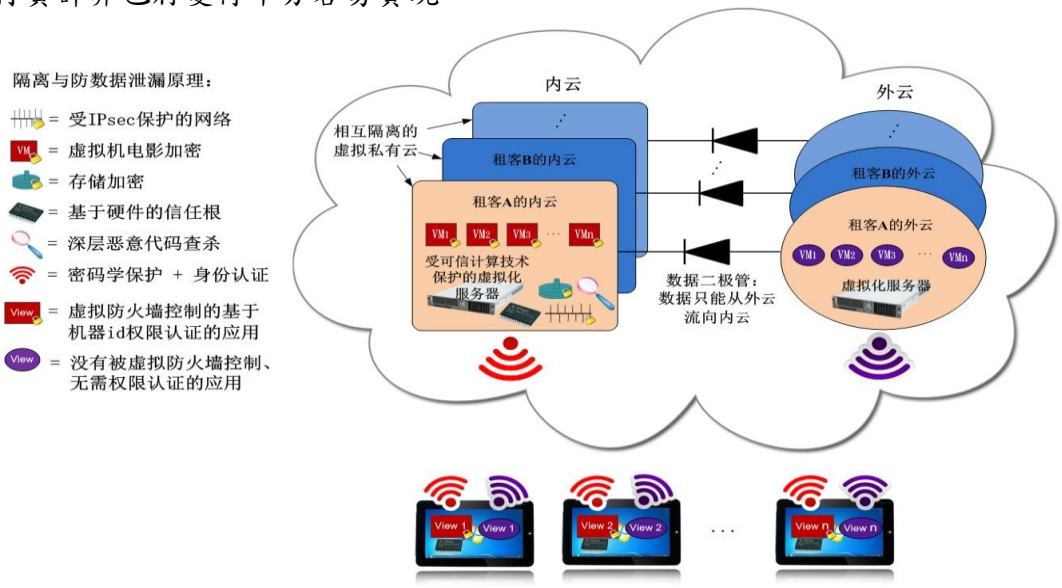


圖 8：NVI 網路虛擬化技術構建的多租賃雲安全解決方案

作爲一種典型應用場景，圖 8 展示了通過 NVI 虛擬化防火牆技術基於虛擬機 ID 來定義組織的外雲、內雲具有單向數據流動拓普構造的 IT 設施解決方案。該方案從網絡通信、數據存儲安全、多租戶隔離到應用隔離等多個方面，爲組織提供了全方位立體化的安全防護措施。該解決方案可以在政務雲、移動辦公雲等多個行業領域中部署。

## 陸、結論

計算虛擬化和存儲虛擬化這兩項技術由於做到了邏輯 CPUs、邏輯磁盤與硬件設備的物理屬性（比如所在位置）脫離關係的徹底性，都已經成功實現了虛擬化，使這兩項技術在工業界得到了廣泛的應用。然而在網絡虛擬化方面，由於對 IT 業務盒子的通信控制手段從來都是集中部署在一些網絡設備上的，因而通信控制始終沒有乾淨地擺脫網絡設備的物理與位置屬性，這個現狀在很大程度上影響了網絡虛擬化技術的發展。

我們觀察到對於在虛擬化架構上運行的 IT 業務盒子，可以利用“在虛擬機這一頭插拔專用網線”（leaf 端節點控制）這一全新手段控制通信。全球分布的虛擬化架構僅需使用一對虛擬機的邏輯身份便可在全球分布範圍執行“網線的插拔”工作，因而可以做到網絡控制與虛擬機的所在位置毫無任何關係。用這種全新的網絡控制手段可以實現徹底的網絡虛擬化，網絡 QoS 工作從此不必再糾纏於處理那些永遠處於變動中的，與網絡設備所在位置緊密耦合的網絡包元數據了，而僅需處理固定不變的虛擬機身份。進一步，我們還利用了密碼學公鑰身份認證原理，實現了在全球範圍安全鎖定虛擬機身份，因而基於處理虛擬機身份的網絡虛擬化技術可以在全球分布部署，實現了可靠、安全、抗分布式拒絕攻擊能力強的雲租戶私有網絡。更進一步，我們還利用了可信計算技術還使密碼學具有基於硬件信任根的保護強度，使雲計算的安全可信度可以向租戶證明。基於密碼學保護的網絡隔離是可信雲服務必須具備的安全質量。

本文提出的網絡虛擬化技術是計算機通信領域裏一種全新的創新嘗試，該技術的逐漸成熟在諸多方面都會涉及標準化制定工作。只有工業界、學術界和研發社區聯手合作，才能完成這項宏大的工程。道裏雲公司將以充分開放的方式與廣大業界同仁廣泛合作，爲網絡虛擬化的技術進步與廣泛應用做出應有的貢獻。

**[誌謝]**

感謝陳克非教授對本文寫作工作的支持以及組織專家評審。道裏雲公司房海峰總體設計了道裏雲安全操作系統，陳山、王文翔對於該系統的實現做出了重要貢獻。

## 參考文獻

[1] DaoliCloud Network Virtualization Infrastructure (NVI) Product Manual, www.daolicloud.com
[2] OpenStack Networking Administration Guide, docs.openstack.org/grizzly/openstack-network/admin/content/
[3] OpenStack Administration Manual home page, docs.openstack.org/openstack-compute/admin/content/
[4] CloudStack, cloudstack.apache.org/
[5] OpenNebula, opennebula.org/
[6] Eucalyptus, open.eucalyptus.com/
[7] Sumayah Alrwais, “Behind the scenes of IaaS implementations”.
[8] OpenFlow Switch Specification Version 1.1.0 Implemented, openflow.org
[9] Open Virtual Switch home page, openvswitch.org/
[10] Nicira, nicira.com/en/openstack
[11] NEC, wiki.openstack.org/wiki/Quantum_NEC_OpenFlow_Plugin
[12] Ryu, osrg.github.io/ryu/doc/using_with_openstack.html
[13] OpenDaylight, wiki.opendaylight.org/view/Main_Page
[14] NSX, www.vmware.com/products/nsx/resources.html
[15] IBM, Kernel Virtual Machine (KVM): KVM Security
[16] Reiner Sailer, Trent Jaeger, Enriquillo Valdez, Ramon Caceres, Ronald Perez, Stefan Berger, John Linwood Grifn and Leendert van Doorn, “Building a MAC-Based Security Architecture for the Xen Open-Source Hypervisor
[17] Reiner Sailer, Enriquillo Valdez, Trent Jaeger, Ronald Perez, Leendert van Doorn, John Linwood Griffin and Stefan Berger, “sHype: Secure Hypervisor Approach to Trusted Virtualized Systems”
[18] Jonathan M. McCune, Trent Jaeger, Stefan Berger, Ramon Caceres and Reiner Sailer, “Shamon: A System for Distributed Mandatory Access Control
[19] Jason Nash, vSphere Security: A Tour of the vSphere vShield Suite
[20] Vmware, VMware vCloud® Networking and Security Overview, Efficient, Agile and Extensible Software-Defined Networks and Security whitepaper
[21] Abhinav Srivastava and Jonathon Giffin, “Tamper-Resistant, Application-Aware Blocking of Malicious Network Connections”
[22] Trip Repoert: Security and Risk Management Community[R]. Garmer Emerging Trends Symposium/ITxpoAptil6-10, 2008
[23] Berger, S. Caceres, R. Goldman, K. “Security for the cloud infrastructure: Trusted virtual

data center implementation". In: IBM Journal of Research and Development
[24] IPv6, en.wikipedia.org/wiki/IPv6
[25] IPsec, en.wikipedia.org/wiki/IPsec
[26] Trusted Computing Group (TCG) Technology, www.trustedcomputinggroup.org

## 附：道裏雲網絡虛擬化架構雲安全操作系統功能

應用道裏雲網絡虛擬化架構 NVI 技術，我們實現了一個雲安全操作系統[1]。該系統可以跨數據中心部署，具有如下性質：

- **靈活的基於虛擬機身份定義的租戶虛擬化私有網絡**
  1. 允許雲租戶組織通過自助服務編輯網頁方式定義構造一個跨數據中心分布的私有"局域"網絡，該租戶私有網絡受到分布式防火牆保護。
  2. 基於虛擬機身份 ID 方式定義的業務網絡，租戶組織可以任意定義互聯或隔離的業務（子）網絡。租戶可以將租用 VMs 的身份列入各個不同的安全組來定義 VMs 之間的通信關係，系統自動將同一安全組內的 VMs 置於同一防火牆內。
  3. 基於虛擬機身份 ID 定義的分布式網絡安全規則，透明支持虛擬機遷移，租戶虛擬機遷移前後所處的任意數據中心物理位置及其 IP 地址的動態變化與虛擬機不變的 ID 無關，因而無需更新按虛擬機身份定義的防火牆策略。
  4. 基於 PKI 公鑰證書認證租戶虛擬機身份的跨多個數據中心的虛擬邏輯網，這是一個具有任意規模的智能邏輯二層。該二層網絡由於自身具備跨物理網絡全球範圍單播（unicast）尋址智能，無需使用傳統的爲通知物理網絡設備組織網絡拓普構造所常用的地址廣播（broadcast）技術，因此邏輯網絡無論多大規模都不會形成廣播風暴。
- **基於密碼學保護力度的安全隔離機制**
  1. 提供對跨多個數據中心的多租戶虛擬網絡控制隔離機制，每個 VM 都有一個 PKI 證書，不僅 VM 的身份可以得到基於密碼學質量的身份認證與完整性保護，而且 VM 的網絡包以及基於快的數據存儲，也都可以被 VMM 透明加密。
  2. 支持外網中的上網行爲控制管理：如僅可瀏覽網頁而不可外發數據，基於網頁白名單/黑名單/網頁內容的網頁瀏覽控制。
  3. 防止存儲系統可能發生數據洩漏的控制措施：透明 I/O 代理技術，在虛擬化基礎設施層對每一個虛擬機的鏡像文件、以及用戶數據磁盤進行實時加解密處理。
  4. 防止網絡系統可能發生數據洩漏的控制措施：透明網絡加密代理技術，在虛擬化基礎設施層對網絡通信報文做實時加解密處理；透明網絡數據流控制管

理技術，在虛擬化基礎設施層對網絡通信報文的源地址與目的地址進行識別與管理，保證通信報文流向符合安全等級保護策略的規定。

5. 基於可信計算平臺 TPM/TCM 模塊的透明密鑰、身份管理、證書管理：每一台物理服務器、虛擬機，每一個用戶、租客都由唯一的身份標識，基於該 ID 採用 TPM / TCM 模塊生成該實體的證書信息。

■ **全面的安全、可靠雲主機管理功能**

1. 租客生命周期管理：創建、刪除、修改等等
2. 加密虛擬機全生命周期管理：創建、啓動、休眠、遷移、關閉、銷毀等等
3. 密鑰系統容災備份、遷移、恢復等等

■ **易用性、透明性、開放性**

1. 用機器身份來定義機器之間的通信關係是個很簡單的工作，無需專業水準的網絡知識與技能
2. 採用符合國際標準的雲計算平臺 Web 前端管理及用戶桌面系統
3. 採用符合用戶使用習慣的前端界面設計，採用先進的 HTML 技術增強系統交互友好性
4. 提供全方位自動化部署工具，支持一鍵安裝模式
5. 採用無需用戶介入的透明密鑰管理方法，完全屏蔽密碼學及密鑰管理的複雜性
6. 透明代理特性：包括透明代理 IO 加解密，透明代理 IPsec 安全通信協議，透明代理數據單向性流動的“數據二極管”功能。由於用戶無法感知這些透明代理的存在，構建在 NVI 架構虛擬化防火牆系統之上的辦公環境與傳統辦公環境的用戶體驗完全相同
7. 開放的系統架構：網絡虛擬化技術及虛擬化防火牆技術可作爲第三方插件方式支持與 OpenStack 系統的集成