國立交通大學
National Chiao Tung University

出國報告（出國類別：學術交流）

赴 University of Minnesota（明尼蘇達大學）參訪 Prof. Pen-Chung Yew 研究團隊

服務機關：資訊工程系
姓名職稱：徐慰中 所長/教授
派赴國家：美國　明尼蘇達大學
出國期間：101/10/23~101/10/30
報告日期：102/07/10

摘要

本計畫主持人曾與 Minnesota 大學的 Pen Chung Yew 教授以及 Antonia Zhai 教授在過去的時間中密切合作，期間共同執行的計畫包含 Agassiz 以及 ADORE。Pen Chung Yew 教授於 2008 至 2011 年間擔任中央研究院資訊科學研究所主任，在此期間，本計畫主持人曾與他就系統虛擬化技術進行密集廣泛的合作，特別是針對 QEMU 系統模擬器中的動態二進制轉譯技術(Dynamic binary translation)。本次計畫主持人訪問 Minnesota，旨在延續與 Yew 教授關於三個面向的研究合作計畫，分別是：1. 強化改善 HQEMU 系統架構與軟體框架，並且針對主從架構模式(Client-Server)環境設定進行調整與優化。2. 針對 HQEMU 系統模式(system mode)模擬的最佳化技術 3. 強化全虛擬化(full-virtualization)系統虛擬機器的效能表現。在與 Antonia Zhai 教授的合作方面，GPGPU 虛擬化技術為其重點所在。Zhai 教授目前於 Intel 進行休假研究 (sabbatical) ，然而她為了參予與本計畫主持人以及 Yew 教授的計畫會議而特別返回 Minneapolis，並且在 Minnesota 大學電腦科學系館進行了為期數天的密集研討。本計畫主持人獲邀參與他們的團隊會議及研討會。歸結而論，本次訪問獲益良多，別具意義。

目次

本文

一、目的

本計畫旨在延續並維持與位在 Twin Cities 的 Minnesota 大學電腦科學及工程學系的合作關係。

Minnesota 大學電腦科學與工程學系共有 38 位教職員，為其 600 位大學部學生與近 400 位研究生提供指導並分享研究經驗。該大學教職員無論在研究以及教學方面皆十分出色。他們在其研究領域深具影響力，所撰寫的軟體以及教科書多被廣泛的使用及引述。他們卓越的表現締造了在研究及教學領域中為數眾多的獎項與榮譽。該大學的教職員同樣地嫻熟於尋求外界的資金挹注，以支持他們的研究計畫。於兩年期間，他們的研究贊助款項已來到 4 億美元，贊助單位涵蓋聯邦政府、州政府以及工業界機構。該大學學系有若干成員榮獲「Presidential Early Career Award for Scientists and Engineers (PECASE)」，並且有 23 位「 CAREER Award」獲獎成員。

Pen Chung Yew 教授任職於資訊科學與工程學系，同時他也是 IEEE Fellow。他曾經擔任多個國際知名研討會主席，包含 ISCA、PACT、HPCA、IPDPS、ICPDS、LCPC、ICS 以及 ICPP。於 2002 至 2005 年間，Yew 教授擔任「 IEEE Transaction on Parallel and Distributed Systems」主編。於 2008 至 2011 年，Yew 教授擔任中央研究院資訊科學研究所主任，並且於 2000 至 2005 年之間，擔任 Minnesota 大學電腦科學與工程學系主任。Pen Chung Yew 教授名列 ISCA 研討會的「hall-of-fame」，該研討會被認為是電腦架構領域方面最具代表性及影響力的指標會議。

Antonia Zhai 教授於 2 年前被晉升為資訊科學與工程學系副教授，她在 speculative thread 的生成與最佳化方面的研究頗具知名度。最近幾年，她的研究專注於 GPGPU 虛擬化技術。

雖然該校的資訊科學與工程學系規模較國立交通大學為小(交大資訊科學與工程學系擁有 70 教職員以及近 1600 位學生)，但他們的系所排名比交大來的突

出。對於我們的教職員而言，這會是一個不錯的起頭機會來與他們洽談合作事宜，以增進加強電腦科學領域的研究工作。

二、過程

本計畫主持人於 10/23 抵達 Minneapolis，並由 Minnesota 大學的 David Du 教授 (該教授在 intelligent network storage systems 相關研究頗具知名度)接機，並接待於其住所。

計畫主持人於 10/24 正式訪問該學系，並由該學系提供辦公室使用。自 10/24 起，本計畫主持人安排了一連串的會議，與 Pen Chung Yew 教授、Antonia Zhai 教授及其各自的研究團隊共商研討。與此同時，本計畫主持人參訪 Pen Chung Yew 教授所開授的虛擬機器(Virtual Machine)課程，以及每週三下午的研討課程。(我將該學系研討會課程詳列於附錄，並且將其中幾項加入交大的研討課程之中。)

在 10/24 當日，我們的討論著重於系統虛擬機的記憶體虛擬化技術(memory virtualization)。更明確的說，我們討論下列兩篇論文：

1. 「Memory Resource Management in VMWare ESX server」, C. A. Waldspurger, OSDI '02: Proceedings of the 5th symposium on Operating systems design and implementation, 2002
2. 「Satori : Enlighted Page Sharing」, G. Milos, et. al., USENIX 2009.

對於系統虛擬機而言，目前的技術瓶頸多半在於實體記憶體(physical memory)，而非處理器。因此，如何管理以及配置、回收虛擬器中的記憶體的相關研究變得至關重要。我們也同時討論了在 Xen hypervisor 中的 ballooning 技術，此技術可被用於回收原先配至給 guest system 的記憶體。

在 10/25 號的討論會聚焦於 QEMU 的系統模式(system mode)強化上。我們討論了一個可能性：是否能夠有足夠的博士班學生來成立一間公司，並且將我們提議的行程虛擬技術(process virtual machine technique)產品化。這可能會帶來甚麼

樣的價值呢？以及我們的技術是否能有足夠的競爭力？

在系統層次的模擬，他的瓶頸發生在模擬記憶體操作指令。由於系統模擬(system emulation)是在模擬客戶端作業系統(guest operating system)，而所有執行中的應用程式中的記憶體操作指令可能會造成 TLB miss 或者分頁錯誤(page fault)。這類的行為將影響實體記憶體的配置方式，以及分頁表(page table)中的內容。這方面與 user mode emulation 不同。在 user mode emulation 中，虛擬記憶體位置(virtual address)會被對應到主機(host machine)中的實體記憶體位置，而 VA(virtual address)與 PA(physical address)間的對應關係是由作業系統來調控。也因此，system mode emulation 必須模擬作業系統的控管分頁以及配置記憶體的行為。這種需求必須要模擬 Soft MMU，而使得模擬的效能大打折扣。在 QEMU 中，我們曾將 user mode simulation 加以最佳化而得到 2 倍到 4 倍的效能增益，然而，對於 system mode simualtion，所做的最佳化技術僅帶來微幅的影響。我們的討論集中於如何有效的應用硬體中的 TLB (Translation lookaside buffer) 來精簡化位址轉譯的程序，而不用經過一連串較為緩慢的過程：VA(virtual address) 轉移到 PA(physical address)，PA 轉移到 RA(real address)。

在 10/26 號的會議則討論到系統虛擬化技術(system virtualization technology)。我們討論了發表於 OSDI 上的一篇論文，研究在虛擬機器中有效的記憶體分享方式。同時，Pen Chung Yew 教授的研究團隊像我們展示了一系列在 Xen、VMware 以及 KVM 上的實作成果。

10/27 是周六，我們並沒有在學校會面，而是參觀了 Mall of America。10/28 周日，我待在 David Du 的家中，閱讀當地新聞報紙以及看電視。

在 10/29，計畫主持人與 Antonia 教授以及他的研究團隊會面，並且討論 GPGPU 虛擬化技術。我們討論了 HSA (Heterogeneous System Architecture) 以及其可能的硬體實作方式、其對編譯器技術的影響等等。HSA 是一個正在演進發展中的開放工業標準，其目的在於支援多樣化的資料平行 (data-parallel) 以及緒程平行 (task-parallel) 程式設計框架模型 (programming model) 。目前，在 HSA 中的成員包含眾多應用處理技開發商，譬如：AMD, ARM, Imagination, MediaTek,

Texas Instrument, Samsung and Qualcomm 等等。HSA 的主要設計目的可歸納於下列幾點：

1.移除 CPU/GPU 的程式設計撰寫中的阻礙

2.減少 CPU/GPU 間資料往返溝通的延遲時間

3.提供對既有程式框架的支援，以期能將此標準套用在更為廣泛的應用中。

4.建立一套基礎規範，作為往後除了 CPU/GPU 之外的其他核心的參考指標

為了支援 HSA 計算，需要一套良好的軟體開發工具。然而，目前尚缺乏一套完整的 HSA 相容的全系統模擬器(full system emulator)。一套全系統模擬器有三點好處：第一，他能夠幫助程式設計者以及執行程式庫開發者來模擬程式執行的狀況，並可在相關硬體還未開發完成之前，預先的進行除錯與測試。第二，他能夠幫助硬體設計者，在正式下線(tap-out)製成晶片以前，預先來評估記憶體架構的設計、硬體排程演算法的設計等等。第三點，他可以被整合進其他的軟體開發工具中，並可用來做系統軟體分析。這是因為全系統模擬器可以運行一套完整的軟體堆疊(software stack)。譬如 Android 系統可以被運行於其中。我們討論了用 QEMU 以及 PQEMU 實做一套模擬器的可能性。我們對於整合 LLVM 編譯器到 QEMU，使得 LLVM 可用來轉譯 GPGPU 裝置指令，有著紮實豐厚的經驗。這項工作將會被分配到交大研究團隊，因為我們曾長時間的研究 PQEMU 以及 LLVM。

10/30 早上，計畫主持人、Pen Chung Yew 教授以及 Antonia 教授齊聚一堂，討論共同向 NSF 提出研究計畫書的相關議題。這提案書將包含：system virtualization, DBT verification, GPGPU virtualization，以及 system mode emulation performance boost 等內容。於此之後，計畫主持人離開 Minneapolis 並返回臺灣。

三、心得及建議

本次訪問程相當豐碩、獲益良多。我們謹慎的將一些研究工作加以分割，好讓他們能夠獨立的分配到各學系來進行。雖然本計畫主持人較為著重在嵌入式系統，且臺灣資訊產業對於嵌入式系統的支援非常充沛，然而，Yew 以及 Antonia 教授更傾向集中心力在企業級系統以及雲端系統，因為此兩項為美國主要研究學群。

Yew 以及 Antonia 教授正在申請 NSF 計畫贊助，如果他們成功獲得補助款，我們的研究也將會有更多的資源挹注，並可吸引更多優秀的研究生投入。並且有機會發現在虛擬化技術方面更多的重要研究議題。

Yew 以及 Antonia 兩位教授非常的強調於訓練他們學生的表達以及演說能力。在每次團體會議中，每位學生被要求準備 10 到 20 分鐘的演講，介紹關於系統設計方面的創意與巧思。計畫主持人以及其他教授並沒有對學生作相關訓練要求。這些訓練對研究生非常重要，我們必須謹慎的看待他。

雖然面談所帶來的效益遠超過遠端通話，但是長途旅行實在是有點累人。並可能會因為時差的關係耽誤了幾天行程來調適。然而，本計畫主持人仍然期望能透過 Skype-meeting 跟兩位教授有常態規律的會議討論。我們曾使用 Skype 以及 Teamview 來完成視訊會議，並且效果優良。

四、附錄

Reading lists from Professor Yew's graduate seminar course:

Process Virtual Machines:
Dynamic Binary Translation and Optimization Systems

Topic 1: Dynamic Binary Optimization (Dynamo)

"Dynamo: a transparent dynamic optimization system", Vasanth Bala, et. al., PLDI '00 Proceedings of the ACM SIGPLAN 2000 conference on Programming language design and implementation, Volume 35 Issue 5, May 2000

"An infrastructure for adaptive dynamic optimization", D. Bruening, et. al, Proceedings International Symposium on Code Generation and Optimization, 2003, CGO 2003.

Topic 2: Dynamic Optimization (Adore)

"Design and Implementation of a Lightweight Dynamic Optimization System", Jiwei Lu, et. al., Journal of Instruction-Level Parallelism 6 (2004)

"Dynamic helper threaded prefetching on the Sun UltraSPARC® CMP processor", Jiwei Lu, et. al., Proceedings. 38th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO-38, 2005

Topic 3: Faster QEMU

"PQEMU: A Parallel System Emulator Based on QEMU", Jiun-Hung Ding, et. al., IEEE International Conference on Parallel and Distributed Systems, ICPADS 2011, Dec., 2011

"HQEMU: A Multi-Threaded and Retargetable Dynamic Binary Translator on Multicores", Ding-Yong Hong, Tenth Annual IEEE/ACM International Symposium on Code Generation and Optimization, (CGO-2012), Apr. 2012

Topic 4: x86 system and process VM

"The Transmeta Code Morphing™ Software: using speculation, recovery, and adaptive retranslation to address real-life challenges", James Dehnert, et. al.,

Proceedings International Symposium on Code Generation and Optimization, 2003, CGO 2003.

“IA-32 Execution Layer: a two-phase dynamic translator designed to support IA-32 applications on Itanium®-based systems”, Leonid Baraz, Proceedings. 36th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO-36, 2003

Topic 5: Dynamic Binary Instrumentation

“Pin: building customized program analysis tools with dynamic instrumentation”, CK Luk, et. al., Proceedings of the 2005 ACM SIGPLAN conference on Programming language design and implementation, 2005

“Valgrind: a framework for heavyweight dynamic binary instrumentation”, Nicholas Nethercote, et. al., Proceedings of the 2007 ACM SIGPLAN conference on Programming language design and implementation, 2007

System Virtualization

Topic 1: Hypervisor on ARM

“Xen on ARM: System Virtualization Using Xen Hypervisor for ARM-Based Secure Mobile Phones”, JY Hwang et., al., 5th IEEE Consumer Communication and Networking Conference, 2008, CCNC 2008.

“The VMware mobile virtualization platform: is that a hypervisor in your pocket?” Ken Barr, et., al., ACM SIGOPS Operating Systems Review, Dec., 2010

Topic 2: VM Migration I

“Fast Transparent Migration for Virtual Machines”, Micheal Nelson, et. al., USENIX’05, 2005

“Optimizing the migration of virtual computers”, C. Saountzakis, et. al., ACM SIGOPS Operating Systems Review - OSDI '02: Proceedings of the 5th symposium on Operating systems design and implementation, 2002

Topic 3: VM Migration II

“Live migration of virtual machines”, C. Clark, et. al., Proceedings of the 2nd

conference on Symposium on Networked Systems Design & Implementation - Volume 2, NSDI’05, 2005

“Live wide-area migration of virtual machines including local persistent state”, Robert Bradford, et. al., Proceedings of the 3rd international conference on Virtual execution environments, VEE’07

Topic 4: Memory Management in System VM

“Memory Resource Management in VMWare ESX server”, C. A. Waldspurger, OSDI '02: Proceedings of the 5th symposium on Operating systems design and implementation, 2002

“Satori: Enlighted Page Sharing”, G. Milos, et. al., USENIX 2009.

Topic 5: Virtual Machine Surveys

“Survey of Virtual Machine research”, RP Goldberg, IEEE Computer, June, 1974

“Virtual Machine Monitors: Current Technology and Future Trends”, M Rosenblum, et. al., IEEE Computer, 2005.

Topic 6: VMM

“kvm: the Linux Virtual Machine Monitor”, A. Kivity, et. al., Proceedings of the Linux Symposium, 2007, page 225-231

“Xen and the art of virtualization”, Paul Barham, et. al., SOSP '03 Proceedings of the nineteenth ACM symposium on Operating systems principles, 2003, Pages 164-177

Topic 7: Xen Enhancements

“Xen and co.: communication-aware CPU scheduling for consolidated xen-based hosting platforms”, S. Govindan, Proceedings of the 3rd international conference on Virtual execution environments, VEE’07, 2007

“Comparison of the three CPU scheduler in Xen”, L. Cherkasova, Performance Evaluation Review, 2007

Topic 8: Memory DeDup

“Difference Engine: Harnessing Memory Redundancy in Virtual Machines”, D. Gupta, et. al., OSDI 2008.

“Decentralized Deduplication in SAN Cluster File Systems”, T. Austin, USENIX 2009

Topic 9: VM Applications

“Terra: a virtual machine-based platform for trusted computing”, Tal Garfinkel, et. al., Proceedings of the nineteenth ACM symposium on Operating systems principles, SOSP’03, 2003

“ReVirt: enabling intrusion analysis through virtual-machine logging and replay”, G. Dunlap, et. al., ACM SIGOPS Operating Systems Review - OSDI '02: Proceedings of the 5th symposium on Operating systems design and implementation, 2002

Topic 10: I/O Virtualization

“Virtualizing I/O Devices on VMware Workstation’s Hosted Virtual Machine Monitor”, J. Sugerman, et. al., USENIX 2001.

“Optimizing Network Virtualization in Xen”, A Menon, USENIX 2006.

Topic 11: Security and Privacy Related Issues

“SubVirt: implementing malware with virtual machines”, S.T King, P. Chen, 2006 IEEE Symposium on Security and Privacy, 2006.

“Managing security of virtual machine images in a cloud environment”, J. Wei, X. Zhang, et al., CCSW '09, Proceedings of the 2009 ACM workshop on Cloud computing security.