本實驗室目前主要的研究方向為深度學習系統優化 (Deep Learning System Optimization), 移動設備 CPU/GPU 合作深度學習, 巨量資料庫 (Database for Bigdata), 雲端計算 (Cloud Computing), 及演算法 (Computer Algorithms). 以下項目均為與中研院資訊所的合作計畫. 所有的計畫均為研究導向, 成果均可發表為碩博士生畢業論文.目前進行中計畫如下.

[Deep Learning System Optimization]
深度學習往往須要耗費大量計算資源及時間，並且應用平行分散式計算。這個計畫的重點在於將平行分散式深度學習加以優化，減少耗費資源。具體的工作項目包括計算結點之間的工作平均分配，使用參數伺服器將深度學習額網路中的權重存取加一優化，例如減少不必要的權重同步，權重的篩選，以及權重的選擇性更新。另外深度學習往往利用大量 GPU 運算，但是 GPU 上的計算資源 (例如記憶體) 有限，往往會對大數量的深度學習權重計算產生瓶頸。這個計畫也針對 GPU 上計算資源有限的現象，提出深度學習各相位計算的排程，使得權重不需額外存副本，藉以減輕 GPU 記憶體的需求。總的來說，這個計畫使用平行分散式計算的優化，提升深度學習的效能。

[CPU/GPU Collaboration Deep Learning on Mobile Device]
隨著移動設備變得越來越普及，客製化移動服務的需求變得越來越重要。如今，移動設備傳感器能夠全天候收集用戶的資料，從而了解他們的使用需求。同時，現代 SoC 的出現使得移動裝置能夠處理機器學習，所以深度學習就成為計算移動裝置上用戶行為的首選方法。不幸的是，訓練深度神經網絡通常被認為計算量過大，不適合移動設備。為了解決這個問題，我們考慮轉移學習，這種技術旨在利用先前學習過的深度學習功能來提高另一個神經網絡的學習性能。我們提出了一個深度學習框架 TransferCL，支持移動設備上的轉移學習。我們的方法依靠多核 CPU 和集成 GPU 的協作來加速移動設備上的深度學習計算。我們的方法同時解決了性能/可移植性權衡，電源效率，和記憶體管理。本計畫同時建議開發一個深度學習框架，以實現深度學習移動設備上的訓練和推理任務。當前最先進的無人監督深度，例如學習技術依賴於生成敵對網絡 (Generative Adversarial Network)。我們的深度學習應用將會實施這樣的架構。

[Cloud-assisted VR]
虛擬實境（Virtual Reality，VR）為了讓使用者能夠有身歷其境的感覺，必須配合使用者的動作（擺頭、移動）在短時間內進行場景及物件繪圖（rendering）運算，以提供高解析度畫面來提高使用者滿意度。但隨著畫面細緻度提高，所需要的運算量亦將大幅提升，很有可能超出穿戴式 VR 裝置本身能提供的運算能力。Cloud-assisted VR 的概念是利用資料中心內伺服器群的資源，來協助處理單一穿戴裝置無法負荷的運算量。當使用者啟動穿戴式裝置時，資料中心會

同步啟動一個對應的 VR 實例（instance），用來負責進行運算。本計畫預計建構 Cloud-assisted VR 框架（framework），負責啟動及處理 VR 實例的資源分配與排程，目標是提供更高品質的 VR 體驗。本計畫目前與中研院資訊所電腦系統實驗室合作中。

[Resource Allocation and Job Scheduling in Enterprise Data Center]
許多企業基於由於方便管理及安全性等理由，會建構自己的私有雲來供公司內部產品及服務使用。但與一般公有雲不同，私有雲的實體資源在大部分情況下是固定的。因此，如何將具有不同性質及要求的工作，分配到合適的伺服器上運行，以減少違反服務協定（Service-Level Agreement，SLA）的數量，便是很重要的問題。本計畫目標為利用資料分析以及機器學習技術，建立應用服務實體資源以及效能模型，依此來幫助企業級資料中心內部資源分配及工作排程。

[Energy-efficient Scheduling on Heterogeneous/Asymmetric Multi-core]
自從 2011 年 ARM 提出 big.LITTLE 架構後，行動裝置及嵌入式系統都開始往異質/非對稱多核心平台發展。異質/非對稱多核心 (Heterogeneous/Asymmetric Multi-core)平台的特點是可以依照工作性質，將需要運算資源的工作放在能提供較高效能的核心運行，而將其他工作保持在運算能力較弱，但消耗電力較少的核心上，以達到省電，延長裝置運行時間的目標。本計畫目標為發展省電排程演算法，在滿足工作要求的情況下，透過排程方式影響使用的核心及時間，以達到節能的目標。

[資料密集型應用程式之平行化技術]
隨著大資料時代的來臨，如何改善資料密集型(data-intensive)應用程式的執行效率，已經成為科學界與工業界所共同關心的研究議題。在此類的應用程式中，資料存取的時間占去大部分程式執行的時間，因此，如何有效率的讀寫資料便是最佳化. 此類應用程式要考慮的首要問題。同時，因為單一伺服器的架構已經無法處理如此巨量的資料，如何利用平行與分散式的架構來處理，也是我們要研究的課題。在此研究計畫中，我們以中華電信的批價系統為主要研究的資料密集型應用。我們有三個主要的方向來改善目前系統的效能：(一) 利用平行處理與分散式架構來避免單一伺服器的效能瓶頸 (二)最佳化資料存取的效率 (三) 以串流處理的方法來即時化使用者的批價資訊。

[分散式圖形資料處理系統]
在海運業中，為了節省貨櫃的運輸與營運成本，海運公司會與出租貨櫃的業者合作，向其租賃貨櫃，以達到充分使用貨櫃的目的，也避免掉購買太多貨櫃而造成使用率低落或是資金的浪費。海運公司為了滿足顧客的需求或是貨櫃使用上的調度，時常會將空的貨櫃由一個港口，運送至另一個港口。然而，這樣的

調度卻增加了額外的運輸成本與能源的浪費。因此，我們提出了一個貨櫃交換的想法，讓不同的海運公司可以藉由交換空的貨櫃來避免掉上述額外的成本。我們將這個想法轉化成，在一般圖中找最大匹配的問題。然而因為貨櫃的數量眾多，我們處理的圖將包含百萬計的節點，使得這個問題無法在單一機器上處理。因此，我們必須建置一個分散式的大型圖處理平台，並且設計有效率的平行匹配算法，來解決這個問題。在這個計畫中，我們將設計一個高延展性的分散式圖形資料庫，包含了一個分散式圖形處理系統和一個分散式圖形儲存系統。我們將發展一系列的技術來優化這些系統，包含了管道式的執行模型，確保本地性的資料分割方法，資料庫分區鎖定，冗餘工作分配，最佳化訊息傳遞以及熱資料預測等方法。我們確信我們的系統能為海運業降低大量的運輸成本，也將對社群網路與物聯網的研究帶來實質的影響。