經濟部 110 年度
《臺灣資安卓越深耕-半導體及資通訊供應鏈資安關鍵技術發展計畫(1/4)》
合作研究計畫

《硬體木馬偵測與防禦計畫》

建議書徵求文件

財團法人資訊工業策進會

中華民國　110 年 3 月 25 日

110年度合作研究計畫建議書徵求文件

一、 簡介

現今物聯網、自駕車等工業技術興起，晶片扮演了不可或缺的角色。若是一晶片或電路無法正常運作，往往牽一髮而動全身，影響整個系統的功能，造成使用者的巨大損失。例如2007年敘利亞的一座雷達未能有效發起空襲警報，起因為系統晶片遭植入後門，導致雷達功能失常；2014年紐約時報揭露美國國家安全局(NSA)在USB埠中植入木馬，使其能夠存取近乎全世界用戶的資料，甚至中國、俄羅斯等國的軍用網路。前述例子再再顯示了晶片安全議題的重要性，不容疏忽。

硬體特洛伊木馬(hardware Trojan horse)泛指晶片中被惡意植入或修改的電路，只有當特定的邏輯閘達到特定訊號條件時會被觸發(triggered)，並且執行惡意安插的載體(payload)電路。其可造成晶片資訊洩漏、功能失效或改變，更甚是直接損壞的後果。而這些觸發條件平時並不容易達成，只有在特殊操作或事件發生時才會滿足。在正常情況下，硬體木馬不影響原電路運作，晶片仍然會得到預期輸出；且硬體木馬並沒有形式化的型態或特徵，每隻硬體木馬都可能是前所未見的型態，這些特性也使得硬體木馬不容易在晶片測試及驗證時被發現。

隨著晶片工業的全球化，設計廠商也越來越趨向「無廠」(fabless)化。晶片在設計與製造的分工更加細緻、專業，卻也讓晶片在不同階段受到各種潛在威脅。在設計階段，使用未認證的第三方工具可能使得元件或參數遭竄改，甚至直接遭植入木馬；在製造量產階段，未被信任(untrusted)的製造商頻繁地接觸、合理地取得晶片設計與製造的細節，亦可能使晶片暴露在風險中。在如此充滿威脅的生產環境中，如何能夠有效率且精準地找出帶有硬體木馬的晶片，是晶片產業所需面對的一大課題。

如同上述的種種情形，若是能建構出一套系統能偵測出該晶片設計中是否含有硬體特洛伊木馬顯得格外重要，由於威脅的種類多元且複雜，歸納出電路是否含有硬體特洛伊木馬的特徵並不是那麼的容易，因此仰賴機器學習的方法解決此問題為其中重要的分支。雖然機器學習在各領域中取得許多輝煌的貢獻，但其不可解讀性是一個龐大的隱憂，這讓許多基於機器學習的系統不被大眾所接受。若是能建構出一套可解讀(interpretable)的機器學習系統，能對於此領域有更進一步的突破。

二、 計畫目標

傳統對於硬體木馬的研究大多以偵測為主，偵測方式大致可分為光學(optical)分析、旁通道(side channel)分析、邏輯測試(logic testing)等，在實際執行上往往需要投入大量時間成本，且因標準樣本(golden model)不易取得且硬體木馬種類與型態複雜，辨識結果不甚理想，這些都是傳統偵測方式所面臨的瓶頸。近年來機器學習技術興起，其中又分為監督式學習、非監督式學習等，也有新興方法如神經網路在崛起，與傳統方法相比能夠有效提升效果，並降低成本，對於這類複雜問題更是一種較為實際的解法。

在硬體木馬領域的研究內容可主要分為兩大面向，一個為硬體木馬的偵測可以將其稱為防守，另一個為硬體木馬的設計與製造可以將其稱為攻擊。本計畫目標為提出機器學習的解決方案，根據需求使用有效的特徵擷取技術取得高品質的訓練特徵，並透過整合資料訓練出準確且高效的偵測方法，最終能提高偵測性能，且優化所耗資源並降低所需時

間；另外希望由可解讀的模型進行建模並基於機器學習的可解釋性分析與生成標籤，目標為得出有用的知識與準則可增強硬體木馬的偵測與設計與製造能力。因為訓練完整後模型內容與判斷準則可被解讀，不只能對模型的內容更加的理解，還能根據模型的內容找出模型的優點與缺點。此外在硬體木馬的偵測部份要能透過模型判別的結果，設計出有效的定位方法，用回推的方式準確抓出硬體木馬實際存在於電路中的準確位置，達到不只能判別還能定位的功能。

另外在硬體木馬的設計與製造相較於軟體漏洞攻擊而言更加嚴重，攻擊者除了需要有特定背景知識外，更需要相當高階的設計手法。因此晶片設計商所面對的威脅對象不止於一般駭客，更可能是具有組織性的公司單位、政府機構等。本計畫將透過前述項目的研究，可得到具有高硬體木馬晶片的辨識準確率(accuracy)的模型，計畫同時透過可解讀(interpretable)的機器學習方法建立硬體木馬的威脅模型(threat model)分析，剖析晶片安全的弱點，找出潛在威脅，針對較弱的部份可以特別生成模型處理不好的樣本，透過重複訓練增強其對硬體木馬的偵測能力，將晶片安全性提升至最高。也可以使用這個特性，使用所得資訊增強硬體木馬不被機器學習模型識別的反偵測能力，生成出隱蔽性更高的硬體木馬，相輔相成的逐步增加模型的偵測的防守與硬體木馬設計與製造的攻擊能力。

三、 計畫範圍

本計畫預計研究以下主要項目：

- 基於機器學習的可解釋性分析與生成標籤
- 可解讀的機器學習技術於硬體木馬的偵測與生成之應用
- 所提出技術與其他方法的比較分析

四、 預期成果

本計畫須配合母計畫需要進行研發，並產出以下成果：

- 於110年06月30日前交付期中報告一篇，並包含機器學習模型之可解釋性分析與生成標籤，以及相關技術之POC。
- 於110年11月30日前交付期末報告一篇，與成果模組一份，並包含原始碼、使用說明以及成果模組之DEMO。

※前述成果如有專利構想或專利申請產出時，需注意專利申請之新穎性(novelty)。因凡經公開發表之研發成果，如擬申請專利，須於公開發表後 6 個月內完成，前述成果如是以論文方式公開發表，將無法取得大陸與歐盟等國之專利。

五、 執行方式

- 合作計畫執行單位應配合本會計畫需求，隨時對計畫細項作調整。
- 於計畫執行期間，合作計畫執行單位須配合計畫所需，不定期與本單位進行研究心得報告與研討，報告內容以計畫範圍相關之技術主題。

六、 計畫期程及預估計畫總經費

計畫執行區間： 110 年 1 月 1 日 至 110 年 12 月 15 日

總經費：800,000元

七、 驗收標準

- 依本建議書徵求文件第四章「預期成果」規定，如期繳交相關成果。

八、 技術能力需求

- 具備硬體木馬、硬體安全研究經驗之研究人員。
- 具備機器學習、深度學習實務經驗之研究人員。
- 熟悉電路模擬工具並具備實際操作或研究經驗的研究人員。