****

**NVIDIA 人工智慧創新研發中心**

**NVIDIA 「Taipei-1 」提案申請書**

申請單位：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

申請日期：113年\_\_\_\_月\_\_\_\_日

1. **「Taipei-1」算力申請說明**
2. 申請資格：
   * 提供我國產官學研及新創進行研發使用；可單獨一家公司/單位，或多家企業/單位聯合申請。
   * 聯合申請需由一家企業/單位主導，並擔任專案主持人。組隊提案主導者必須為具法令地位法人(legal entity)，且須羅列使用者名單。
   * 申請資格為「非商業用途」：禁止使用「Taipei-1」運算資源產生營收 (例如雲端服務、挖礦等)，但可用來開發模型或解決方案做為研發或未來產品或服務。
   * 申請者應**已備有訓練所需資料**，且**須具算力使用經驗**。
3. 提案重點：
   * 徵案題目包括但不限於：生成式AI相關應用、大型語言模型、數位孿生等。
4. 專案運算資源與使用流程：
   * DGX運算資源：16台DGX H100，每台8張H100 GPU。
   * OVX運算資源：8台OVX 2.0，每台8張L40 GPU。
   * 每專案使用之時間共6周：第1周on-boarding作業包含申請帳號、資料上傳、教育訓練等前置作業，第2至5周運算使用共計4周(含假日)，第6周off-boarding 如資料下載、解除帳號、刪除資料等。
5. 申請期間：
   * 自公告日期至113年5月30日下午5:00截止，請將申請書 email 至產業技術司(執行團隊)窗口：

* 工研院產科國際所出小姐 電郵：minghui\_chu@itri.org.tw / 電話：03-5916637
  + 申請使用期間：113年7月1日至9月20日，每專案以6周為限(詳申請書)。

1. 審查機制：

* 經濟部產業技術司籌組專家委員會，針對提案申請書進行審議，分就DGX運算群、OVX運算群進行審查及核定。
* 計畫採批次收件、批次審查，依最終評核結果及推薦順序，免費提供Taipei-1算力，申請通過者將公告於A+企業創新研發淬鍊計畫網站（https://aiip.tdp.org.tw/）。

1. 審查重點：

* 主題：技術前瞻性、商業化潛力。
* 算力需求：獨立或聯合多家企業/單位，可充分運用6週專案算力資源為優先。
* 申請者過往經歷：具備訓練資料、大算力使用經驗。

1. 專案時程規劃：

* 4月19日：公開說明會
* 5月30日截止收件：申請者提交申請書(電郵方式：出小姐minghui\_chu@itri.org.tw)
* 6月上旬：完成書面資格審查
* 6月中旬：進行會議審查(通知通過資格審查之申請者到場簡報說明)
* 6月下旬：公告通過名單
* 專案啟動：

預計第一批(113年7月1日到113年8月9日)

預計第二批(113年8月12日到113年9月20日)

1. 權利義務

* 每個專案皆應於使用運算資源前與NVIDIA簽署保密及使用者規範等相關協議。
* Taipei-1使用者需遵循Taipei-1用於測試、評估和開發，用戶不應將專有或敏感資訊帶入該環境。

1. **「Taipei-1」使用申請書**

| **項目** | **申請內容** |
| --- | --- |
| 主導企業或單位/專案主持人 | 請勾選：   1. □ 單獨申請 2. □ 聯合申請   主導企業或單位：  專案主持人姓名：  職稱：  電話：  電子郵件信箱： |
| 聯合組隊申請之其他企業或單位 | 共有\_\_\_企業/單位，請羅列如下： |
| 專案主要聯絡人 | 姓名：  職稱：  電話：  電子郵件信箱： |
| 執行AI、數位孿生專案經驗說明 | (例：用OO模型與資料去train，並達成甚麼效果) |
| 專案重點說明   1. 專案主題：包括以下題目但不限於如生成式AI應用、大型語言模型、數位孿生等(註明有無其他政府計畫補助) 2. 專案的重要性：整體之研究規劃與重要性，本次專案之目的及解決那些問題 3. 預期產出：專案計畫預計成果 4. 完成專案後的效益，如我國提昇前瞻技術、帶動產業共創、提升商業化價值、培育人才等 5. 專案對於大算力之需求理由 6. 專案的執行方式說明及可能遭遇的困難 | 1. 專案主題： 2. 專案的重要性： 3. 預期產出 4. 預期效益： 5. 算力需求說明： 6. 執行方法與困難： |
| 專案時程 | 請勾選：  □ 第一批7月1日~8月9日  □ 第二批8月12日~9月20日  □ 第一批或第二批皆可 |
| 專案運算資源 | 請勾選(擇一)：   1. □ DGX運算群；所需 \_\_\_台(上限16台) 2. □ OVX運算群；所需 \_\_\_台(上限8台)     所需儲存空間(以GB或TB為單位)： |
| 資料集描述 |  |
| 所需之NVIDIA軟體及軟體開發套件  備註：   1. 申請DGX運算群者請說明計畫使用DGX H100時將需要的NGC軟體內容 (<https://ngc.nvidia.com/>) 2. 申請OVX運算群者請說明計畫使用OVX時將需要的NVIDIA Omniverse Enterprise Applications軟體內容 (<https://docs.omniverse.nvidia.com/>) |  |
| 計畫參與使用人員  備註：  請提供人員清單，包括企業或單位、姓名、職稱、email、電話。 |  |
| 補充資訊(請以附件表示) |  |

**補充說明**

1. Taipei-1 系統說明

Taipei-1是一座NVIDIA在台灣投資建置的人工智慧與數位孿生運算超級電腦，目的是為了支持台灣NVIDIA人工智慧創新研發中心(以下本文簡稱為本中心)研發所需的運算資源。整體運算資源的部分運算力預計將投入與台灣進行與本中心目標一致的共創或研究計劃。

整體系統包含2個部份的主要運算資源，一個部分是主要運用在AI運算與NVIDIA最新的前瞻技術開發，採用NVIDIA DGX H100運算主機構成。另一個部分主要運用在Digital Twins 數位孿生與 NVIDIA 最新的前瞻技術開發，採用NVIDIA OVX運算主機構成。相關系統說明如下：

1. DGX 系統

整體依據NVIDIA DGX SuperPOD (<https://www.nvidia.com/zh-tw/data-center/dgxsuperpod/>)參考架構進行規劃建置，採用64台NVIDIA最新人工智慧超級電腦DGX H100 (<https://www.nvidia.com/zh-tw/data-center/dgx-h100/>)，每台 DGX H100 主機具備8個NVIDIA H100 GPU透過最新一代的NVLink以及NVSwitch技術進行整合，整體共計512個H100 GPU。DGX主機內每一GPU彼此之間具備900GB/s的雙向溝通頻寬，主機與主機之間透過8條最新InfiniBand技術搭配NVIDIA Quantum-2 InfiniBand (400Gb/s 高頻寬與低延遲高速網路技術<https://www.nvidia.com/zh-tw/networking/quantum2/>)進行主機間 GPU 運算溝通，為運用在AI運算與NVIDIA最新的前瞻技術開發的超級電腦。

1. OVX 系統：

整體架構依據NVIDIA OVX SuperPOD (<https://www.nvidia.com/zh-tw/data-center/products/ovx/>)參考架構進行規劃建置，採用32台OVX 2.0 主機進行整合，每台主機搭配8個NVIDIA L40 GPU，整體共計256個L40 GPU，透過NVIDIA Spectrum X (<https://www.nvidia.com/zh-tw/networking/spectrumx/>)高速乙太網路架構進行整合，投入符合實際物理情況的Digital Twins數位孿生和生成式人工智慧，包含AI訓練及推論，NVIDIA OVX™ 系統可以提供最佳高度擬真符合實際物理的3D繪圖技術和運算效能，以及在數位孿生應用研發中整合人工智慧運算的各項技術。

本專案分享DGX 系統、OVX 系統各25%算力，故為16台DGX H100、8台OVX 2.0 。

1. Taipei-1 環境軟體說明

DGX 系統中將使用NGC(NVIDIA GPU Cloud) 相關最新與最佳化的AI 開發套件 (<https://ngc.nvidia.com/>)，包含 TensorFlow Framework、PyTorch Framework、 Computer Vision、Conversional AI、NeMo Framework for Generative AI、BioNeMo healthcare for Generative AI、TAO for pre-training models and optimization 以及 Triton Inference Server 等等 AI 開發環境與套件 (詳細請參考 <https://developer.nvidia.com/>)以協助合作單位與企業使用最適切的開發環境進行研究或共創。

使用者介面搭配 NVIDIA Base Command Platform MLOps 使用介面提供服務，使用手冊請詳閱<https://docs.nvidia.com/base-command-platform/index.html>。

OVX 環境中搭配 NVIDIA Omniverse Enterprise 軟體，整合Omniverse Nucleus server、USD Composer、USD Presenter 以及 Audio2Face、Kaolin、Machinima 以及 Isaac Sim 相關 Applications 套件，幫助研發人員共創Digital Twins數位孿生的相關研 NVIDIA Taipei-1 使用申請書 Page 6 of 8 究目標，相關Applications完整文件與資料請參閱 <https://docs.omniverse.nvidia.com/> 。

1. 專案主題

包含但不限於下列主題：

* 生成式AI (Generative AI) 相關應用：包含但不限於生成式AI於音樂、圖像、金 融、法律等領域的應用。
* 大型語言模型：大型語言模型相關技術於語音處理與自然語言處理的開發與研究，以及大型語言模型技術在各種實際場景的應用。
* 數位孿生：建立3D應用程式、3D設計協作、數位孿生、機械人模擬、合成資料生成 等技術在各產業或研究領域的應用。

其他NVIDIA 軟體開發工具套件（SDK）相關主題（請參考以下備註）：利用NVIDIA SDK 提供的技術和工具所開發更多創新應用，以需要大規模運算者為佳。 以下舉例介紹數個NVIDIA SDK 做為參考：

* NVIDIA NeMo框架（[https://www.nvidia.com/en-us/ai-data science/generative-ai/nemo-framework/](https://www.nvidia.com/en-us/ai-data%20science/generative-ai/nemo-framework/)）：該框架專注於語音處理與自然語言處 理的眾多方向，以滿足各種實際應用場景的需求。
* NVIDIA BioNeMo（<https://www.nvidia.com/en-us/gpu-cloud/bionemo/>）：該平台結合深度學習技術，用於藥物分子設計、蛋白質結構預測等方面，促進生物技術 與藥物研發的交叉合作。
* NVIDIA Modulus（<https://developer.nvidia.com/modulus>）：開發物理資訊神經網絡（Physics-Informed Neural Network, PINN）應用於科學計算。
* NVIDIA CUDA Quantum（<https://developer.nvidia.com/cuda-quantum>）：協助研究人員在一般計算機上模擬量子系統。
* NVIDIA CLARA（<https://developer.nvidia.com/industries/healthcare>）：醫療相關的研究與應用，用以幫助提升診斷和治療效果。