AI-Stack

AI-STACK ENTERPRISE

機器學習協作管理平台

Version 4.27.0

使用者操作手冊

2025-09-26

目錄

1. 機器學習協作管理平台簡介

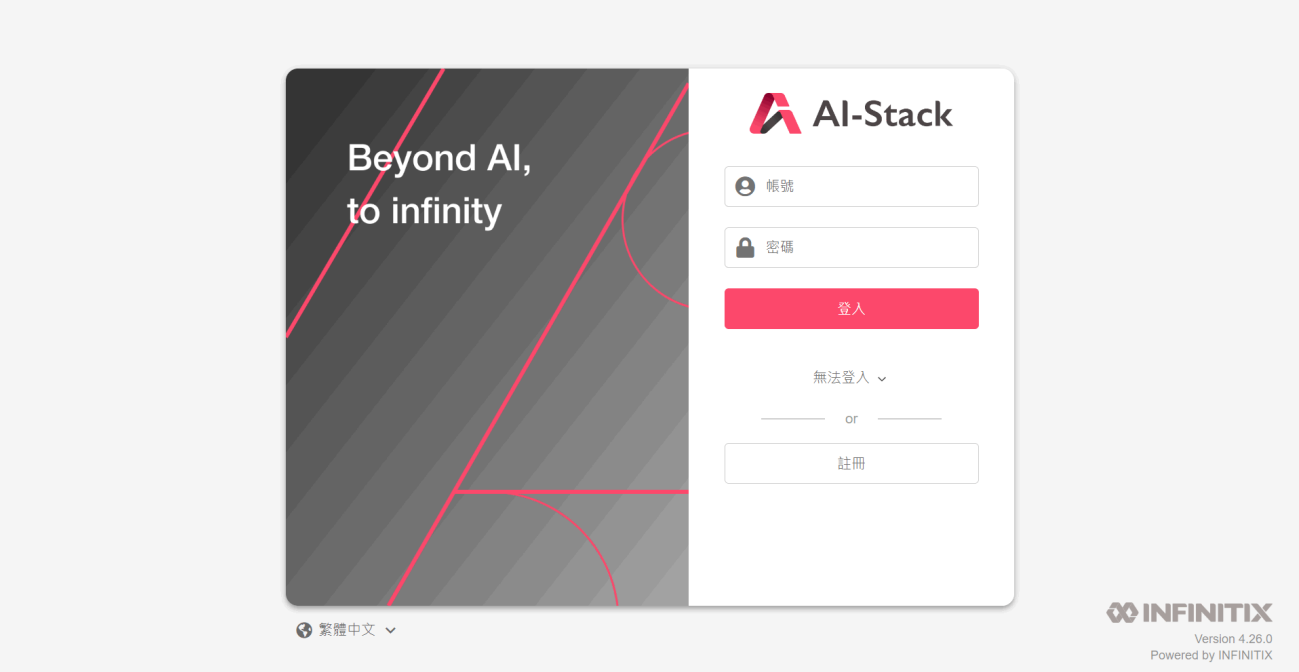
本手冊針對使用者所撰寫，將介紹 AI-Stack 機器學習協作管理平台之功能及操作方式，使用瀏覽器版本建議為 Chrome 60 或以上 / Firefox 60 或以上，瀏覽器若舊於上述版本將無法順利操作本系統。

本系統提供使用者自助服務的環境，能按需求自由配置容器資源，選取並掛載所需的 CPU、GPU、記憶體、網路磁碟機及 AI 機器學習框架（如 TensorFlow），並可取得其他與工作相關的系統資源資訊。

本系統採用專案為基礎（Project-based）之設計原則，讓使用者可依特定計畫目標、部門別、或小型實驗、內部競賽、教育訓練、活動展示等多種使用情境，建立專案項目進行區分，並針對不同專案配置所需資源，且個別使用者可同時參與多個專案，讓資源之利用、預算之管控得以有效管理。

2. 登入平台

下圖為使用者登入頁面，請在此輸入使用者帳號與密碼進行登入。

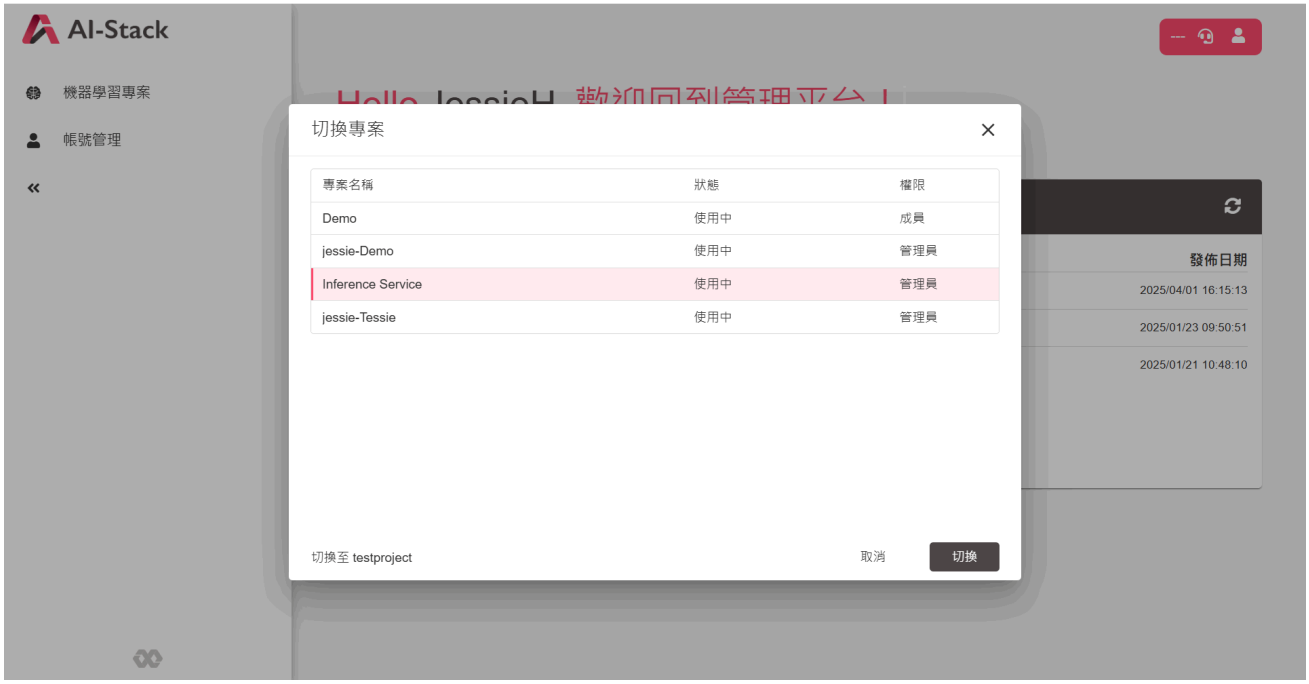


首次登入平台，若平台資源尚未開通完成，畫面會顯示【雲平台設置】頁面。開通相關操作請參考：4. 開通服務一節。

3. 平台首頁

點擊右上角  圖示，即可檢視目前登入使用者名稱，欲離開本系統則點擊 [登出] 即可，從下拉式選單可選擇系統顯示語系。

點擊右上角  圖示，可開啟專案清單，並且可檢視在專案中之身份是成員或是管理員。開始操作本系統前請確認已選至欲操作的專案，避免誤用不同用途的專案資源。



4. 開通服務

【雲平台設置】頁面會顯示使用者在本系統資源平台的開通狀況。當開通狀況顯示為[開通中]時，請稍待流程執行完成，如下圖。

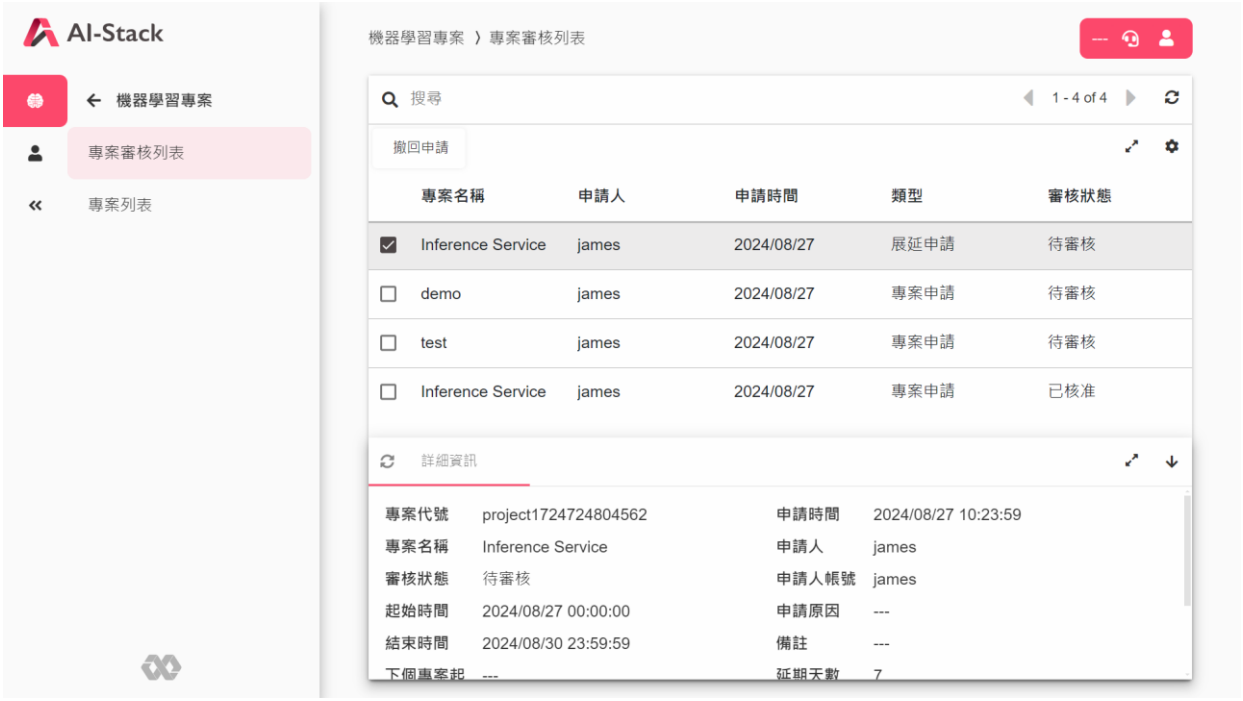
點擊  重新整理後，若開通狀況為[已開通]時，即可開始使用機器學習服務。

5. 機器學習專案

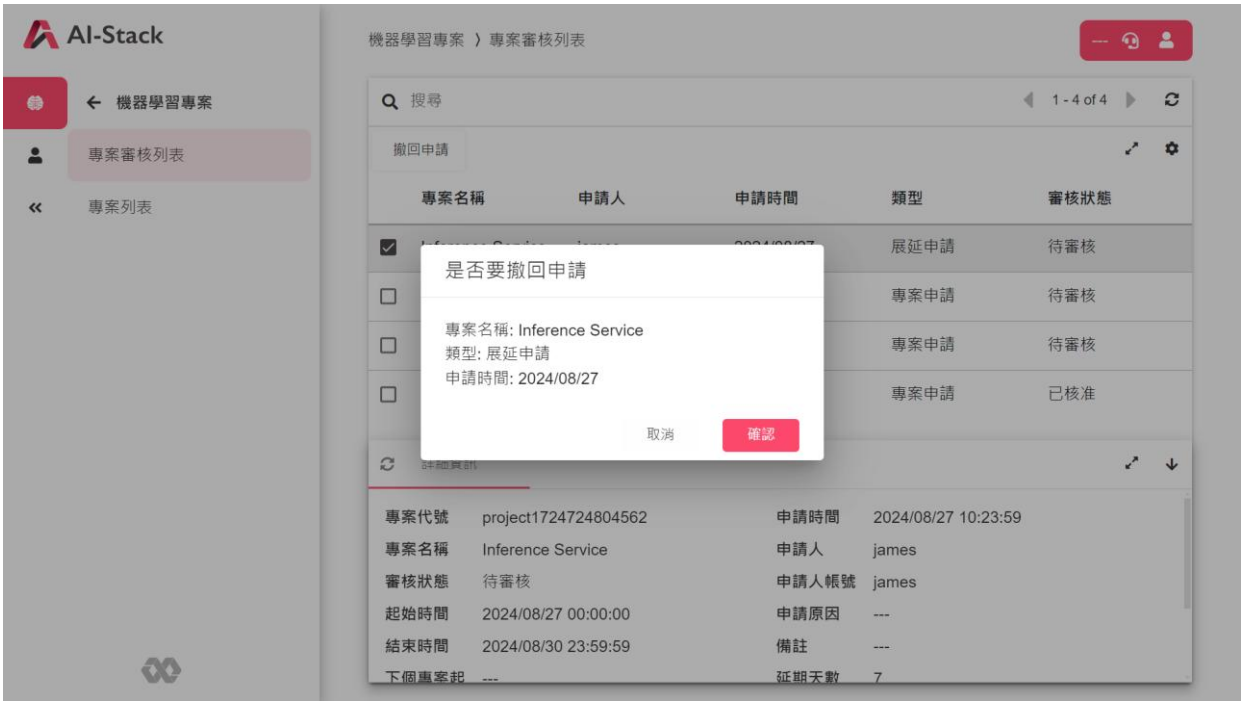
使用者可透過專案列表頁面，以卡片檢視各個專案的基本資訊與資源使用狀況。其中需注意事項為，專案有起始日與結束日，當專案結束日到期前會使用郵件通知該專案管理員進行專案展延，若專案到期後無展延，則系統會自動刪除專案內容器進行資源釋放。

5.1 專案審核列表

供使用者查看已送出的申請單，包含專案申請、展延申請、額度申請。



若申請單狀態為 [待審核] 則提供撤回申請功能。



5.2 專案列表

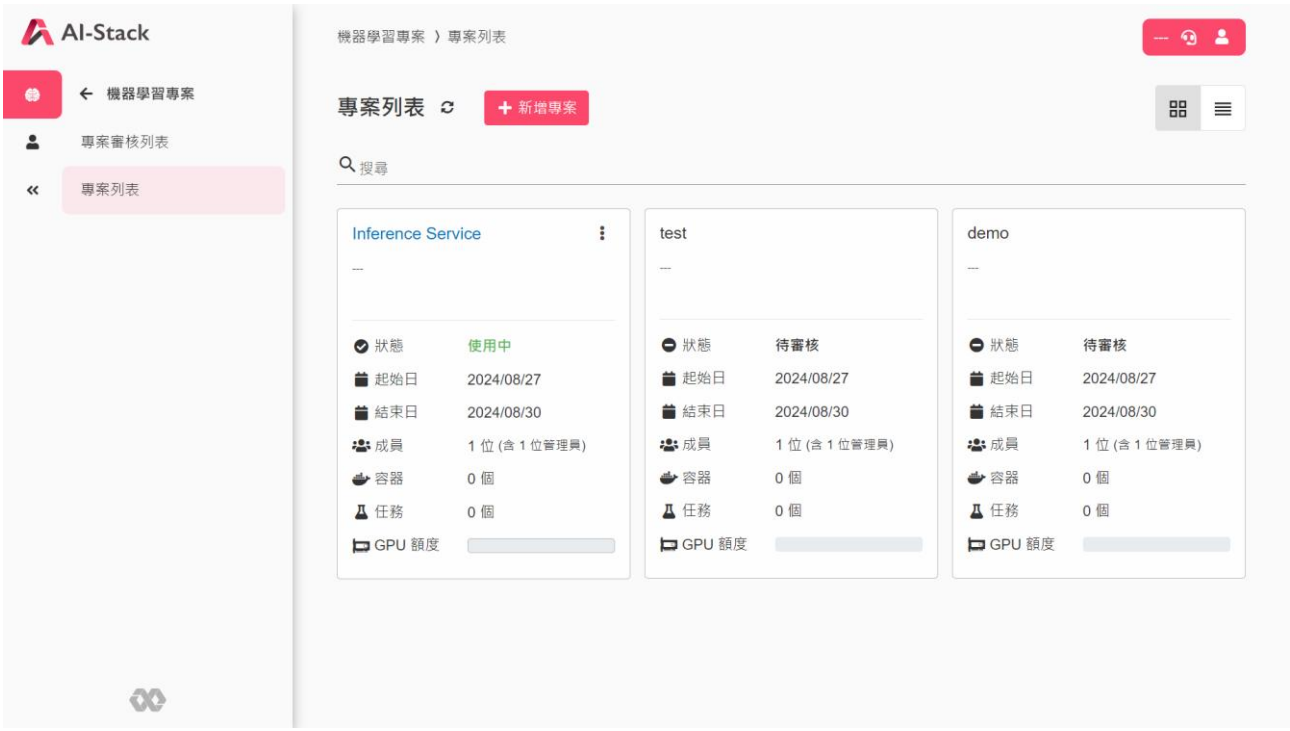
- 專案狀態

  專案依據開始、結束時間與審核通過與否有以下幾種狀態：

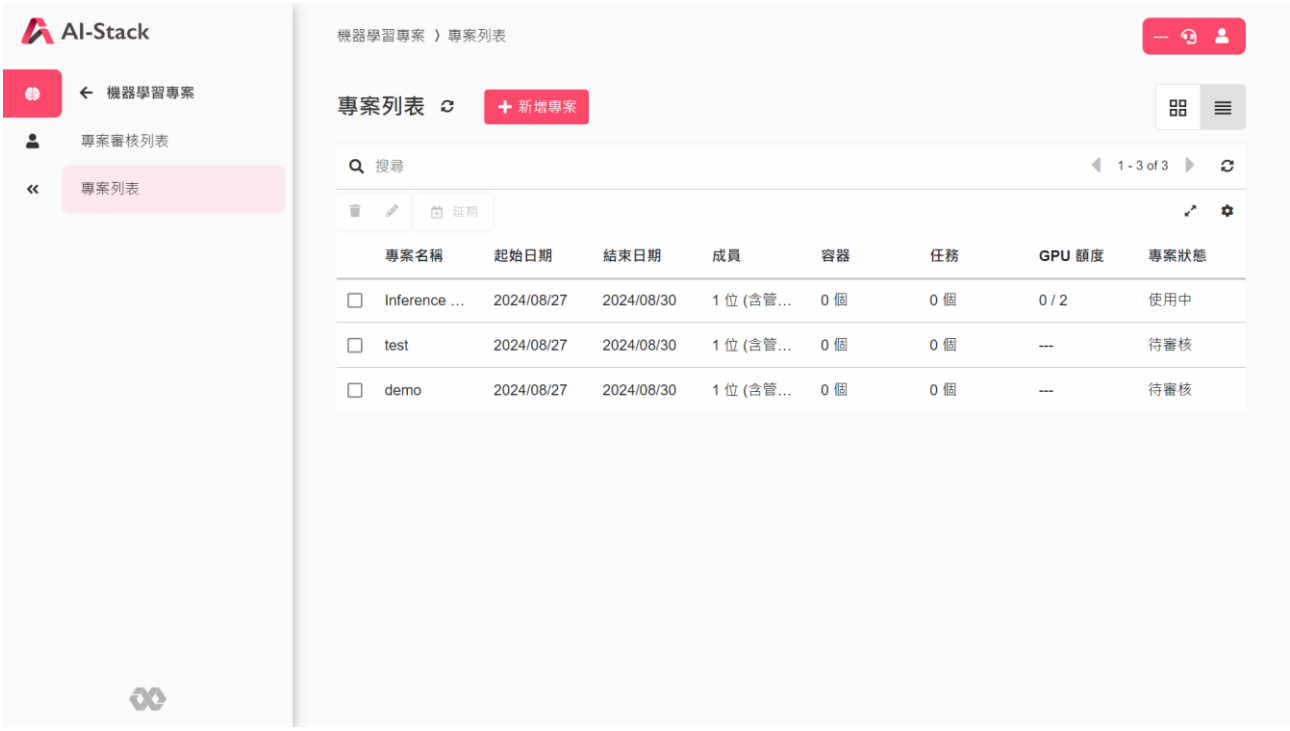
狀態	說明
待審核	需等待平台管理者核准。
使用中	可正常進入專案操作各項功能。
已核准	平台管理者已核准專案申請，但尚未到專案起始日。
已駁回	平台管理者駁回專案申請，駁回原因可於該專案申請之專案管理者信箱中查看通知信。
緩衝期	當專案結束日期已到，且未申請展延。此時進入 14 天緩衝期，系統將自動刪除專案內所有容器，釋出佔用資源，並於緩衝期結束後，自動刪除專案。

- 查看總覽資訊

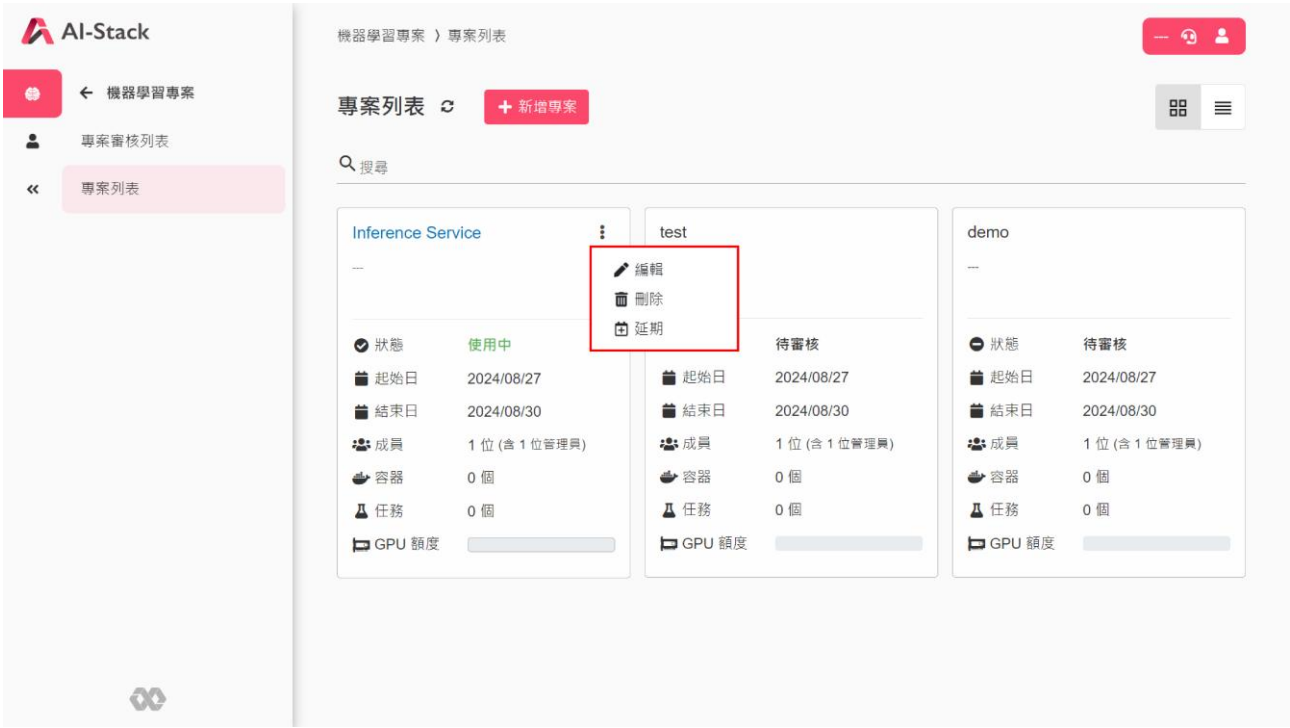
  點選【專案列表】功能選單後，預設將以卡片模式檢視各個專案的基本資訊及資源使用狀況，其中專案狀態 [待審核] 為建立專案送出申請後之初始狀態，需待平台管理者核准始可開始操作。



本功能亦提供切換不同的專案列表呈現方式，點擊  圖示可切換為列表模式。



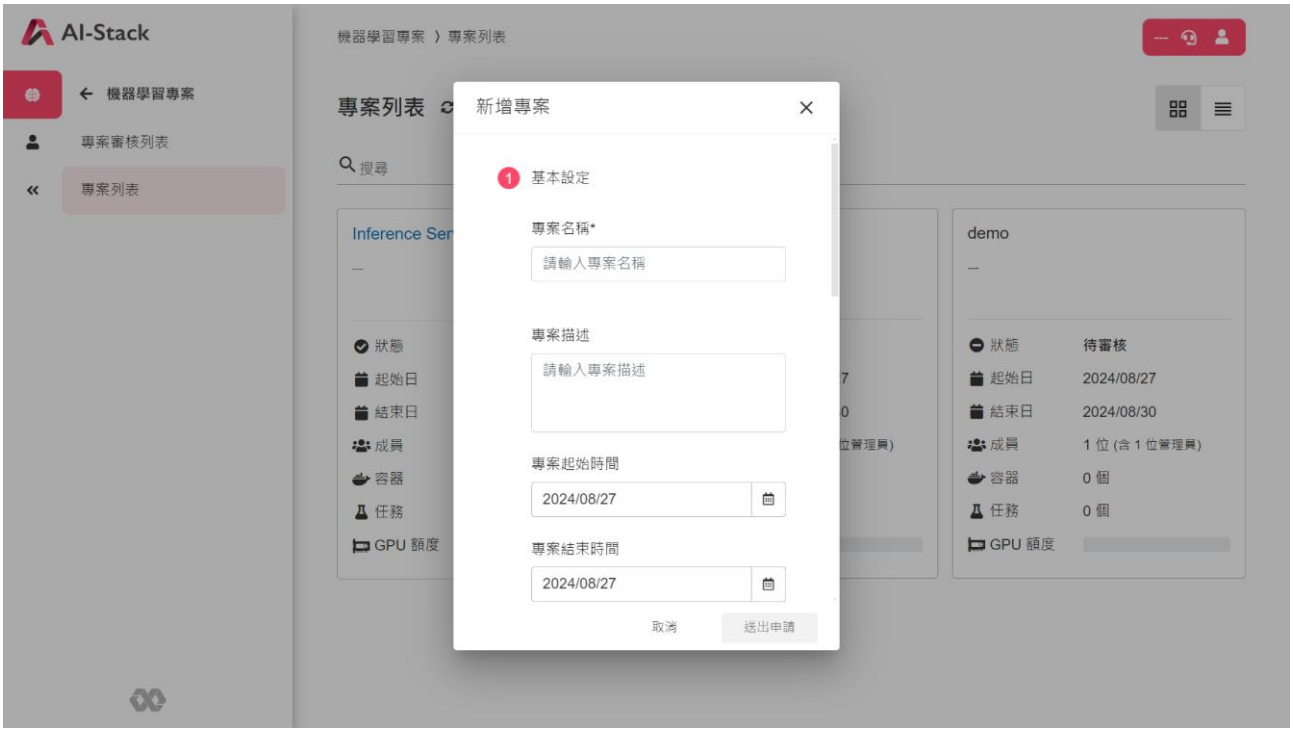
在卡片模式下，在欲編輯的專案卡片點擊右上方之 ⋮ 圖示後，將出現編輯、刪除、延期等功能選單，此功能僅限於該專案之管理員。



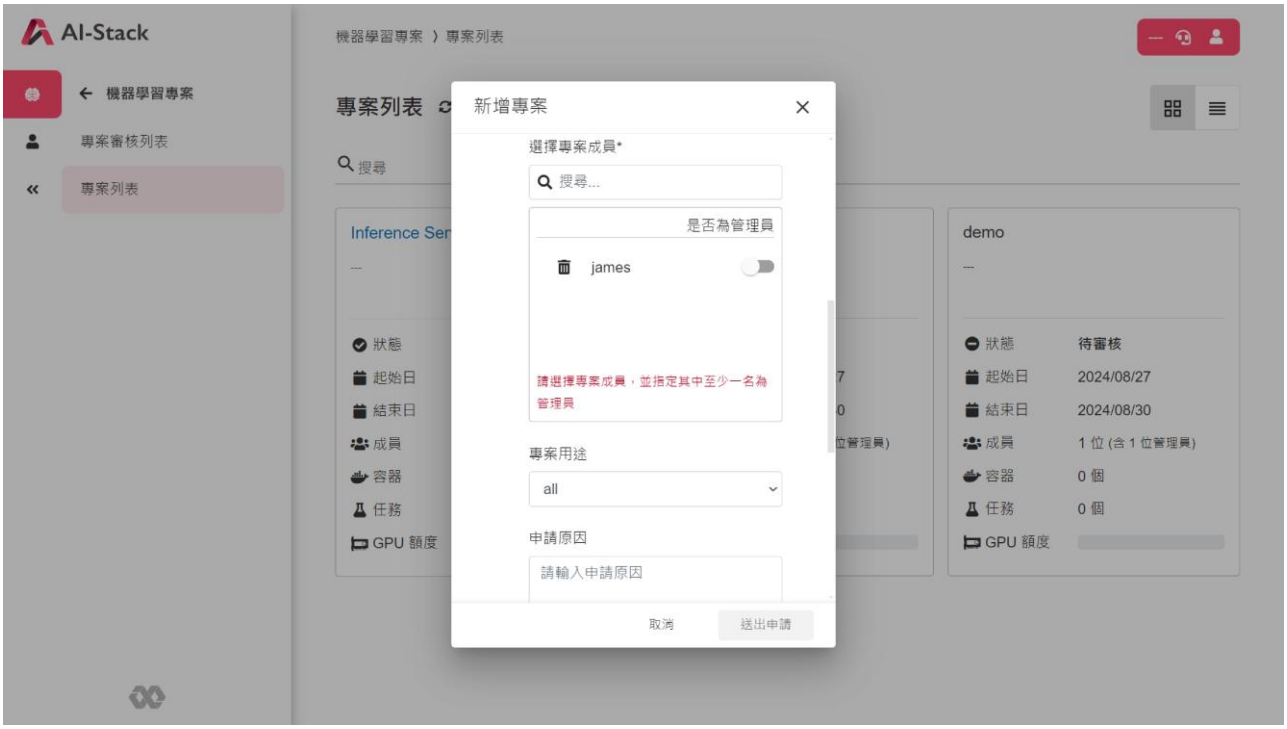
- 新增專案

  供新增專案之申請，送出後需由平台管理者審核通過始得建立容器等操作。平台管理員核准後，申請者將收到專案審核結果通知郵件。

  - 專案名稱：輸入可辨識之專案命名。
  - 專案起始與結束時間：可依實際需要填入起始與結束時間。



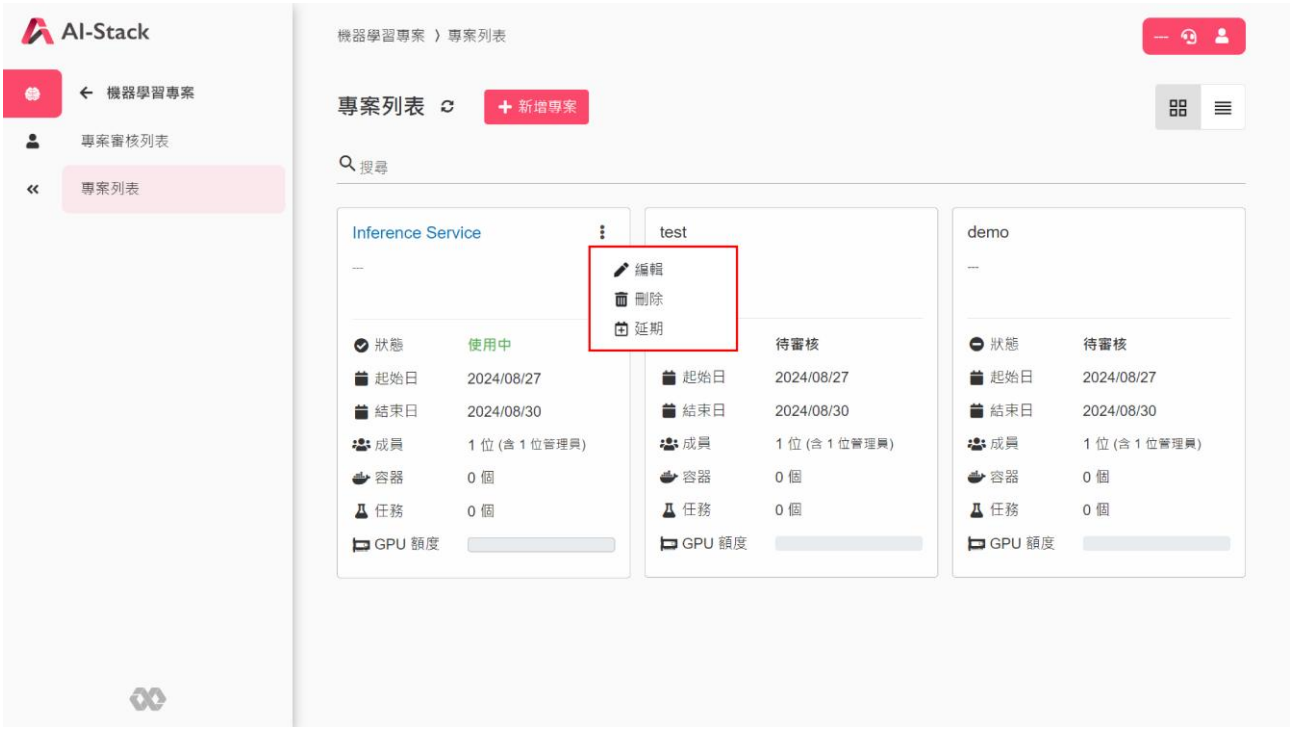
- 選擇專案成員：新增專案時需指定最少一名專案成員。
- 專案用途：根據本專案目的進行用途選項指定（將影響可用資源）。



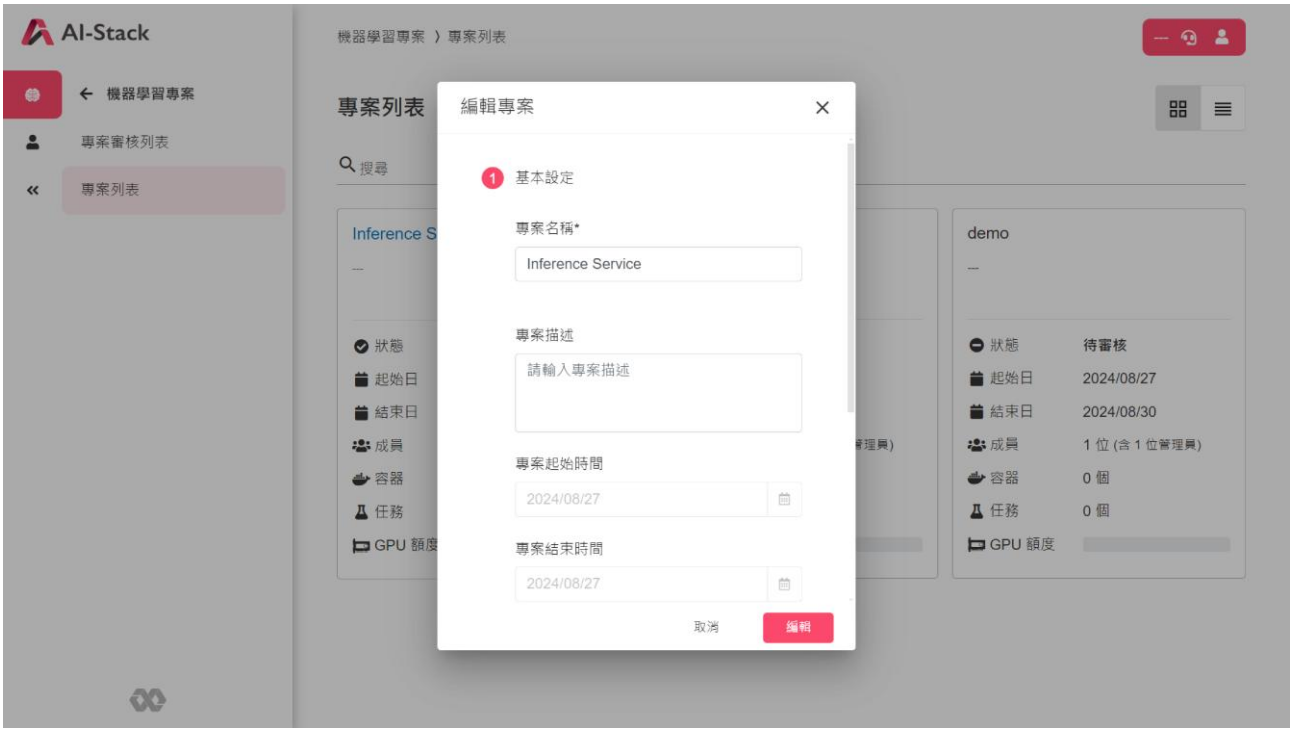
平台管理員核准後，申請者將收到專案審核結果通知郵件，如下圖內容。

- 編輯專案

  點選欲編輯的專案卡片，並點擊右上方 ⋮ 圖示，按下編輯，即可進入 [編輯專案] 頁面，僅限於專案管理員可進行編輯。

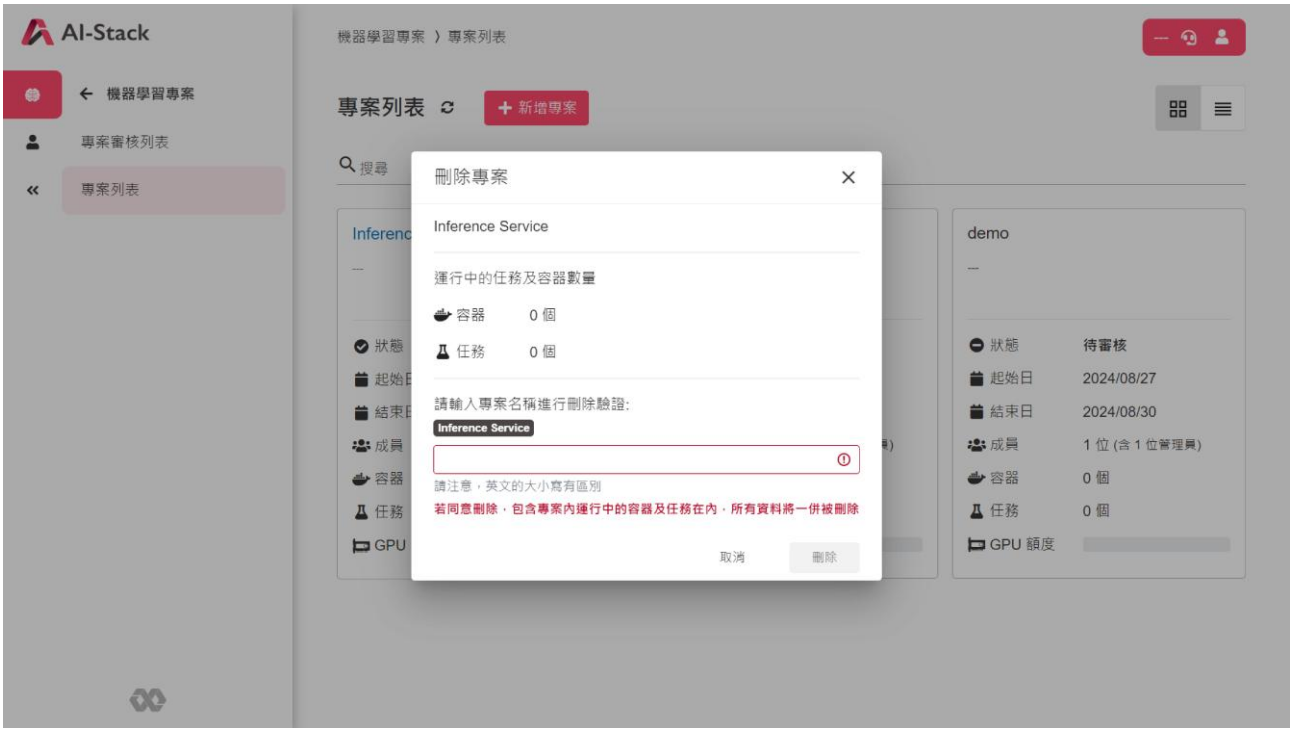


編輯專案可以修改專案名稱、專案描述，並可以新增專案成員與指派為管理者。



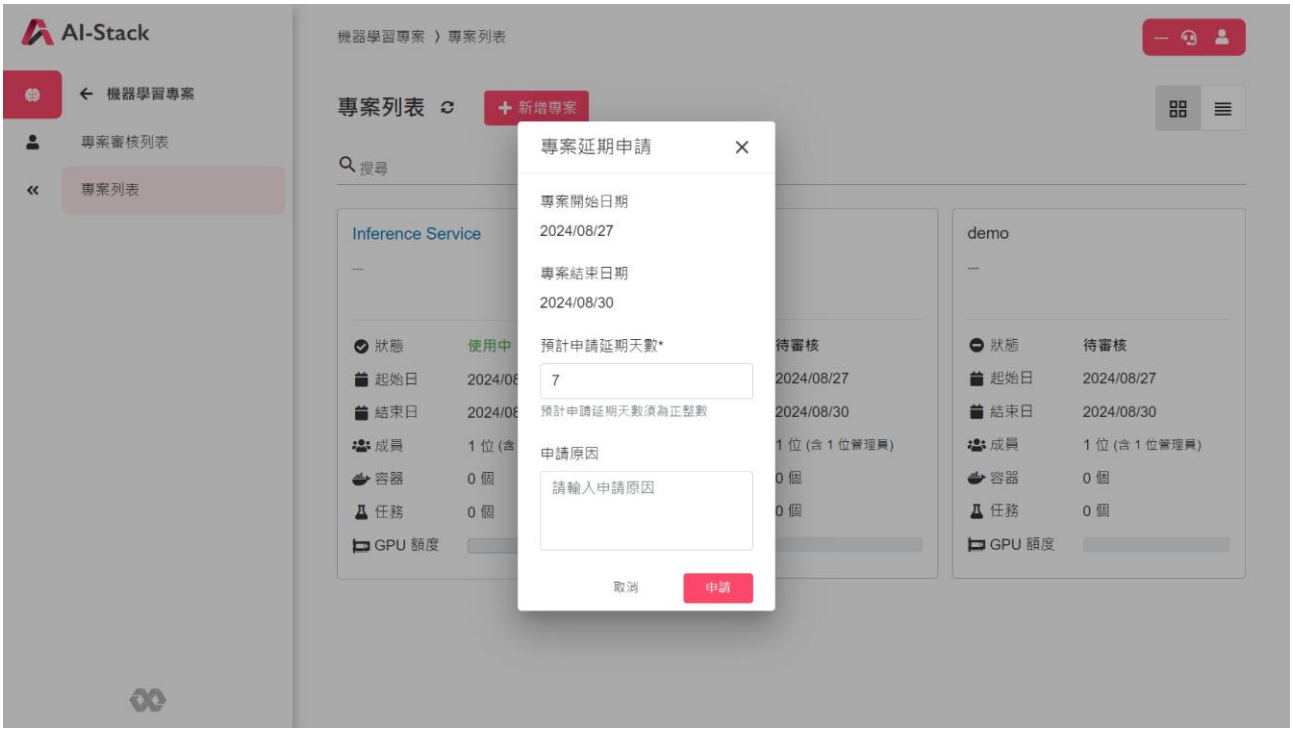
- 刪除專案

  刪除專案時會連同專案中的容器與任務一併刪除，為避免誤動作，需輸入畫面的刪除驗證碼才會予以刪除，此功能僅限於該專案管理員。

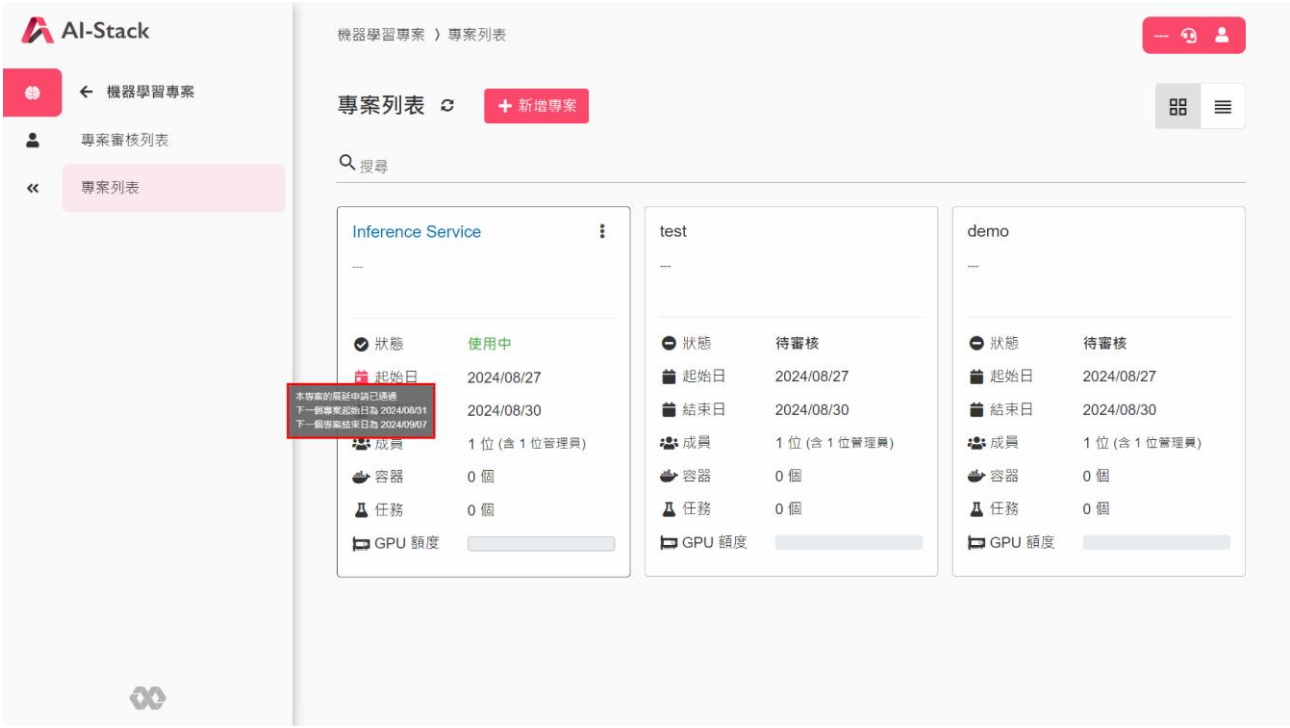


- 專案延期

  在欲申請延期的專案卡片點擊右上方 ⋮ 圖示，再按下延期，即可進入「專案延期申請」畫面，填入欲申請的延期天數即可送出延期申請。

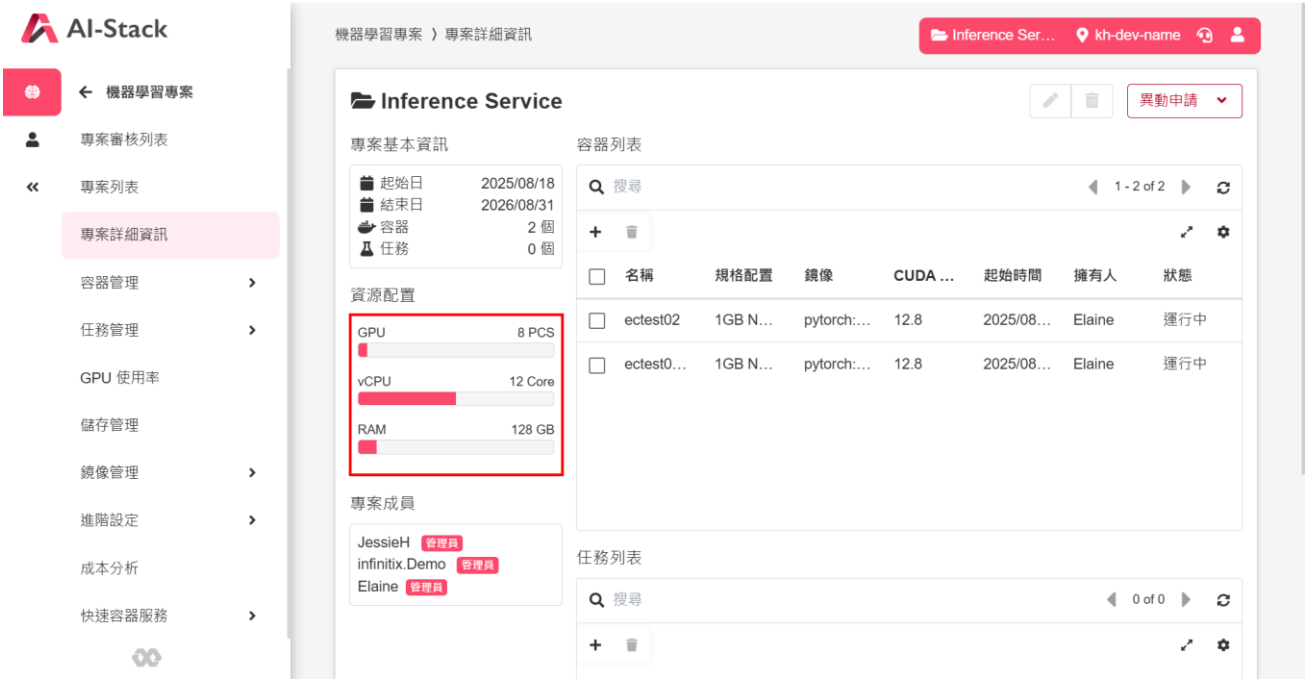


專案延期申請核准後，原本的專案起始日期的圖示會變成紅色，點選圖示可以出現下一個專案的開始與結束日期。



5.3 專案詳細資訊

在【專案列表】點擊要查看的專案即可進入【專案詳細資訊】，可於容器列表中查看所有專案成員的容器 * 、該專案資源額度限制、專案成員與目前執行的機器學習任務。使用者也可以於專案詳細資訊中的左上方資訊檢視該專案的起始與結束日期、容器及任務總數。

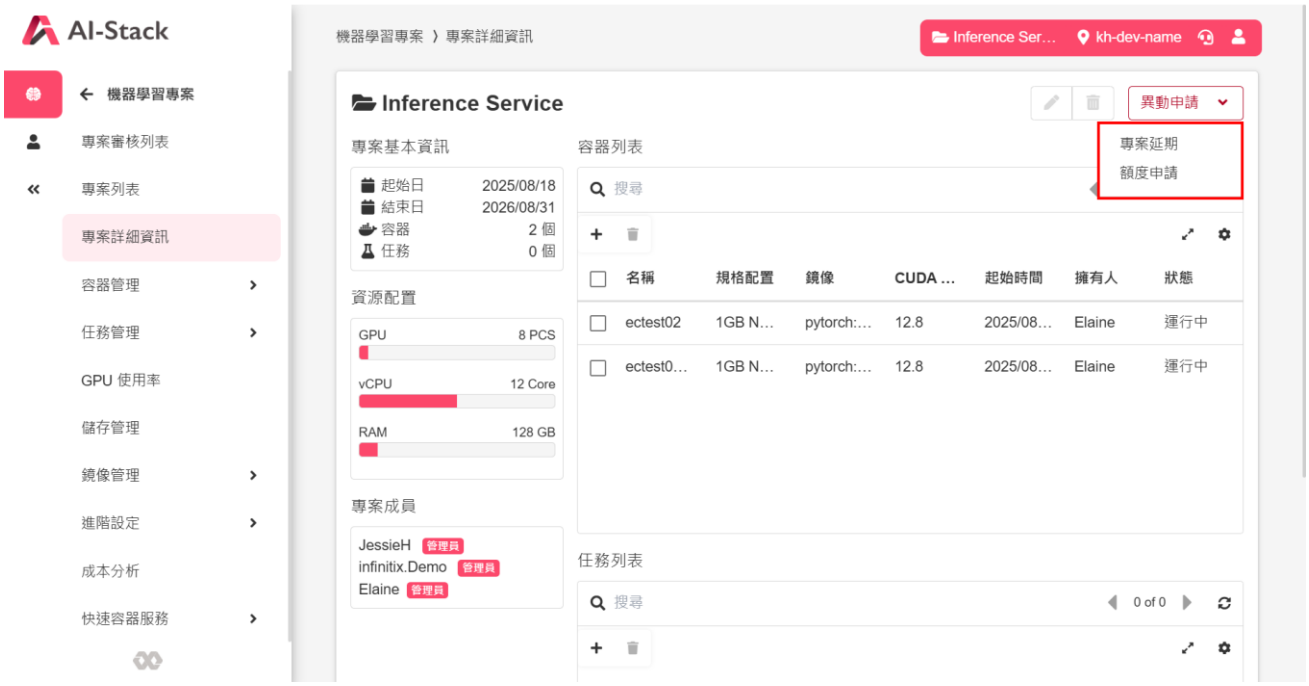


紅框處 [資源配置] 為呈現該專案的 GPU、vCPU 和 RAM 的額度與剩餘可用資源。

* 備註：預設值為專案管理者可查看所有成員容器，一般成員僅能查看自己的容器。

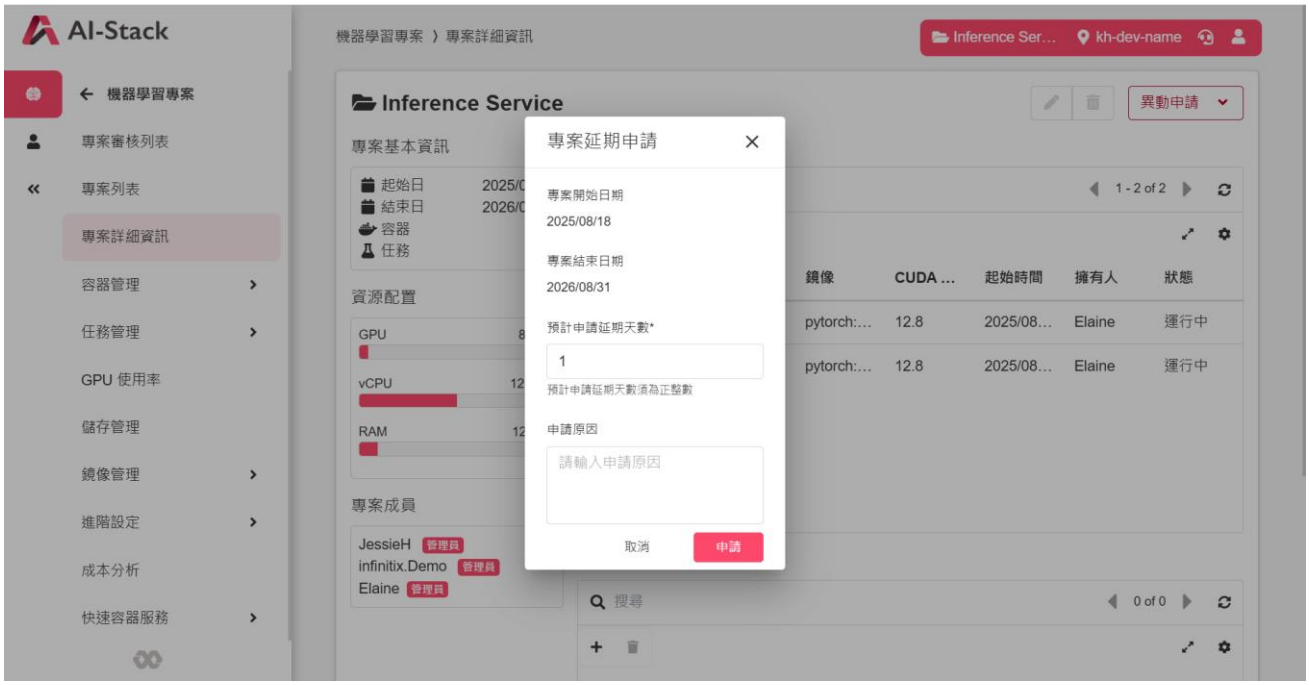
- 專案異動申請

  專案異動申請提供專案延期與額度申請兩種功能，其中專案延期申請為該專案管理員權限。



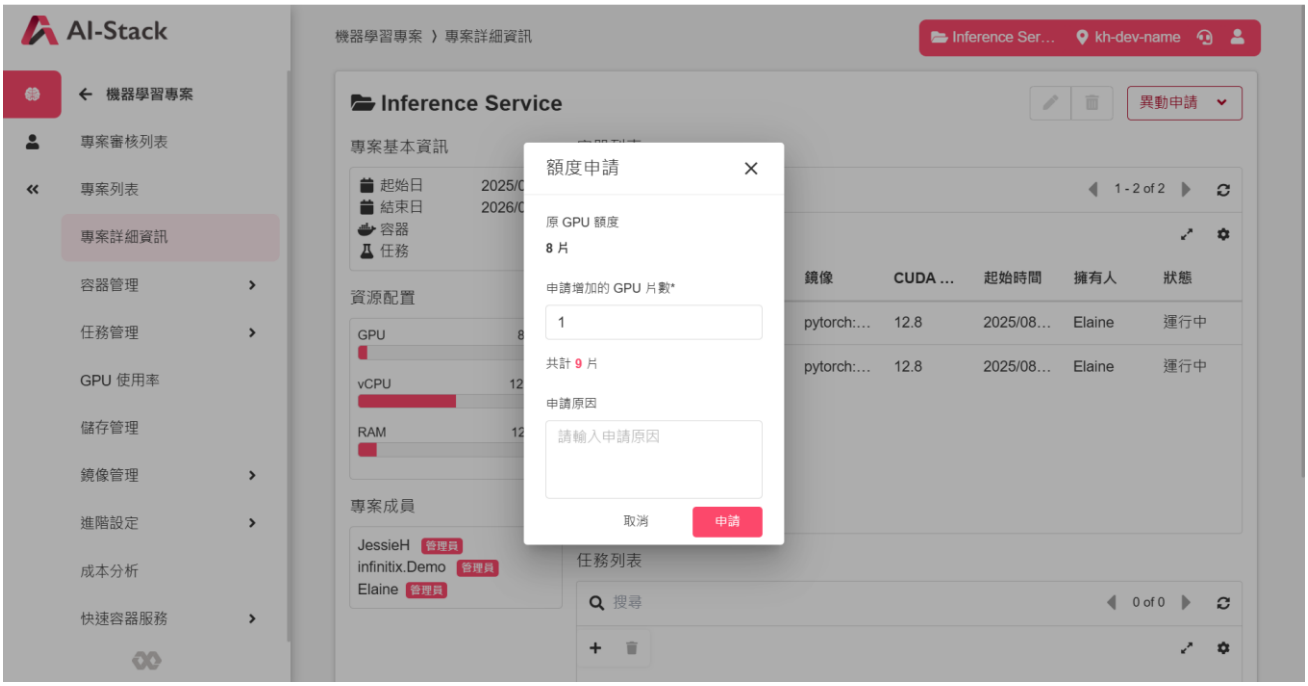
  ■ 專案延期申請

  僅專案管理員可申請專案延期，送出申請後需由平台管理者核可方可延期。



■ 額度申請

額度申請會依據現有 GPU 額度再加上申請的額度，此申請單送出後需由平台管理者核可才會進行額度調整。



● 專案到期通知

當專案結束日期已到，若無進行延期申請，系統會自動寄送專案到期通知信，此時將進入 14 天緩衝期，緩衝期內無法建立任何資源。

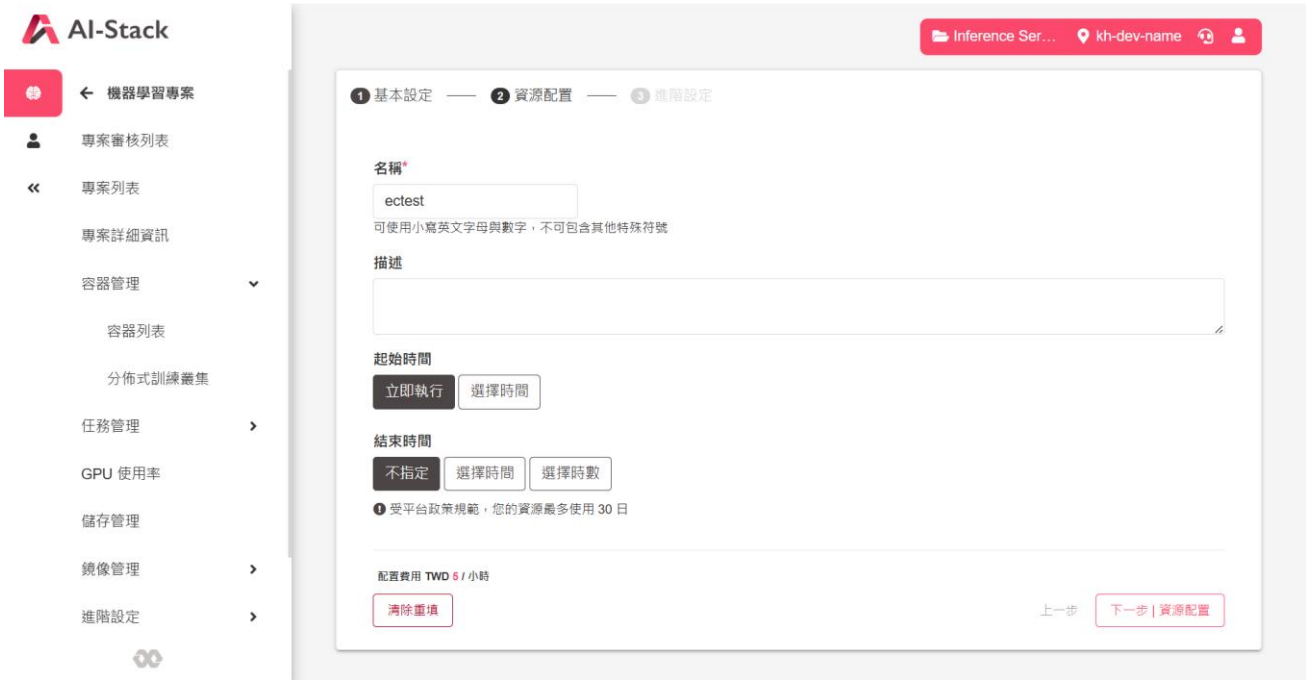
5.4 容器管理

本系統提供網頁操作讓使用者快速、簡易地使用機器學習服務，並依需求自由選擇 AI 機器學習框架（如 TensorFlow、PyTorch 等）。

5.4.1 建立容器

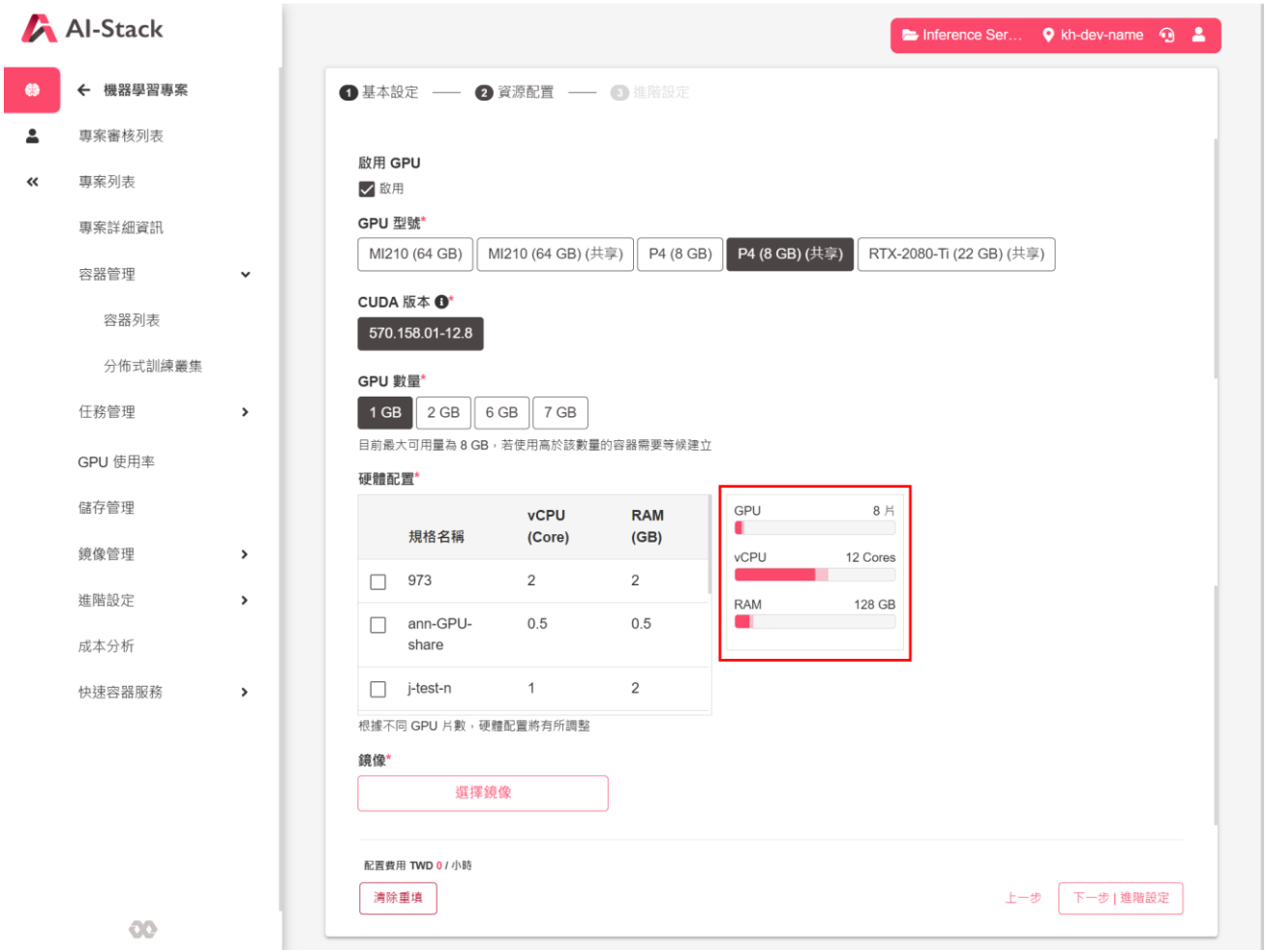
5.4.1.1 一般情形

- 進入【容器列表】頁面並按新增 ⊞。
- 輸入名稱。
- 設定起始時間、結束時間。
  - 起始時間：選擇容器要立即執行或選擇時間。立即執行指的是送出後立即建立，亦可透過選擇時間指定特定日期建立。
  - 結束時間：選擇容器是否要指定終止時間，若選擇不指定則可無限期使用；選擇時間則於設定的指定時間刪除；選擇時數則於容器建立後設定的時數後刪除。
- 確認內容後，點擊 [下一步 | 資源配置]。

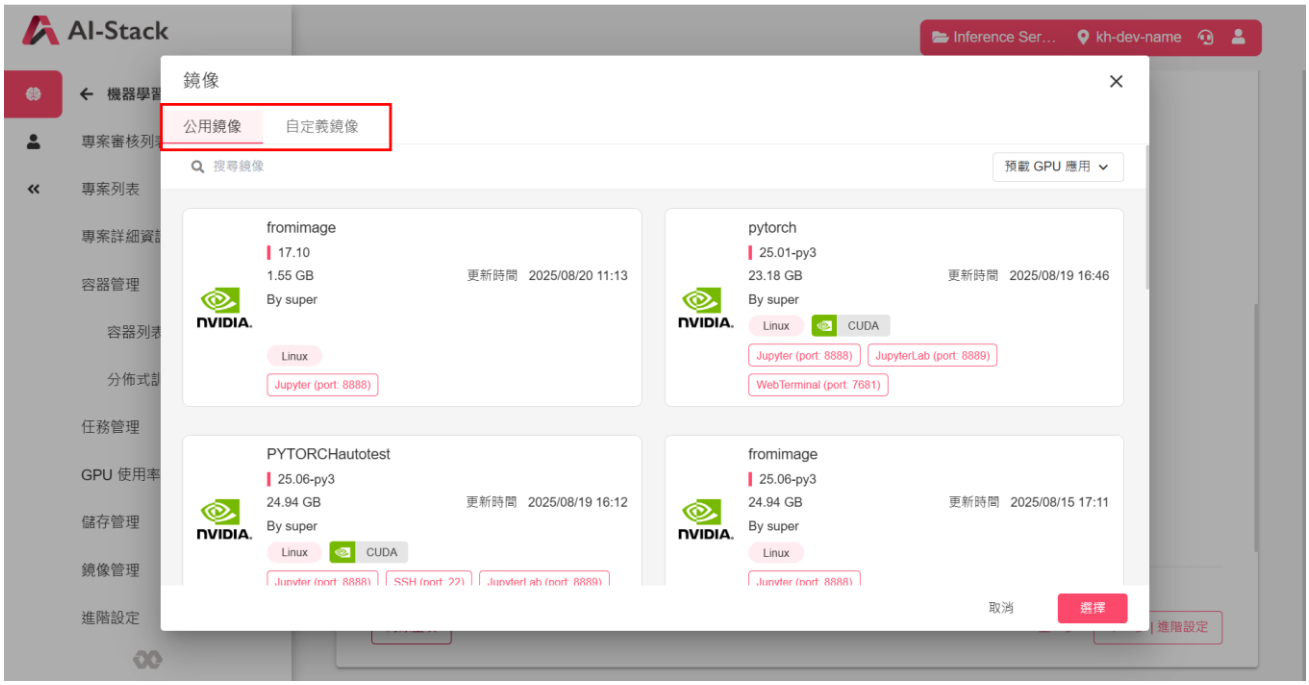


- 選擇欲使用的 GPU 型號、GPU 數量（已綁定後台設定）。
- 選擇欲使用的 CUDA 版本。
- 選擇欲使用的硬體配置，包含 CPU 核心數、記憶體。

  硬體配置右方圖示（下圖紅框處）為可用資源提示，會根據所選擇的規格變化。

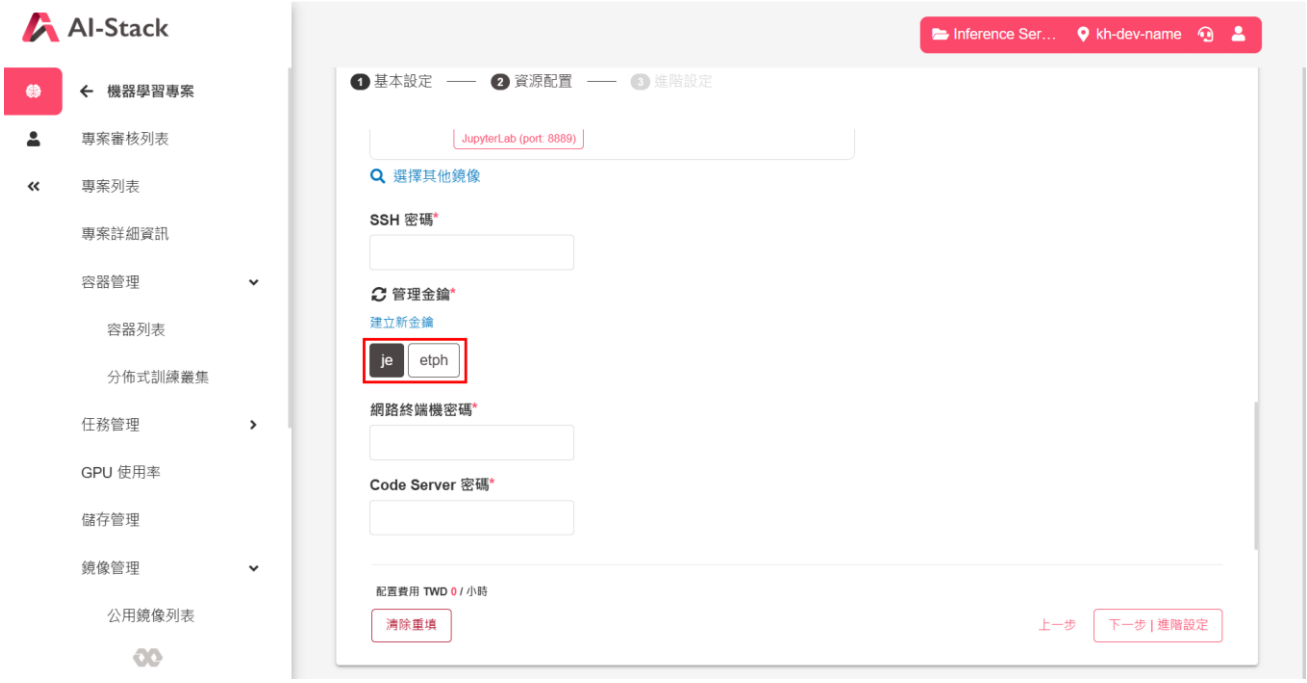


- 鏡像：分為公用鏡像或自定義鏡像，確認欲使用的鏡像後點擊 [選擇]。

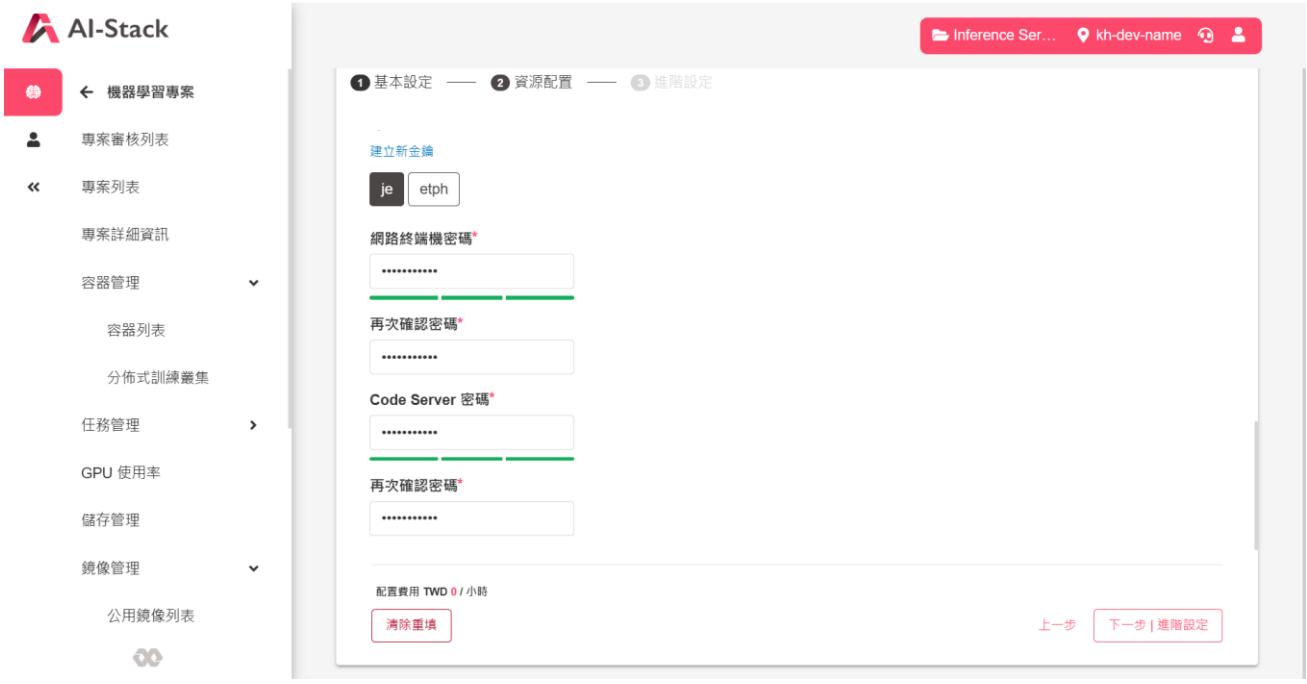


- 管理金鑰

  當使用者曾在平台上建立金鑰且所選鏡像模板之 SSH 登入方式開啟 [透過金鑰] 選項時，可選擇使用哪一組金鑰連入。



- 若需要輸入密碼，確認全部填入且內容無誤後，點擊 [下一步 | 進階設定]。



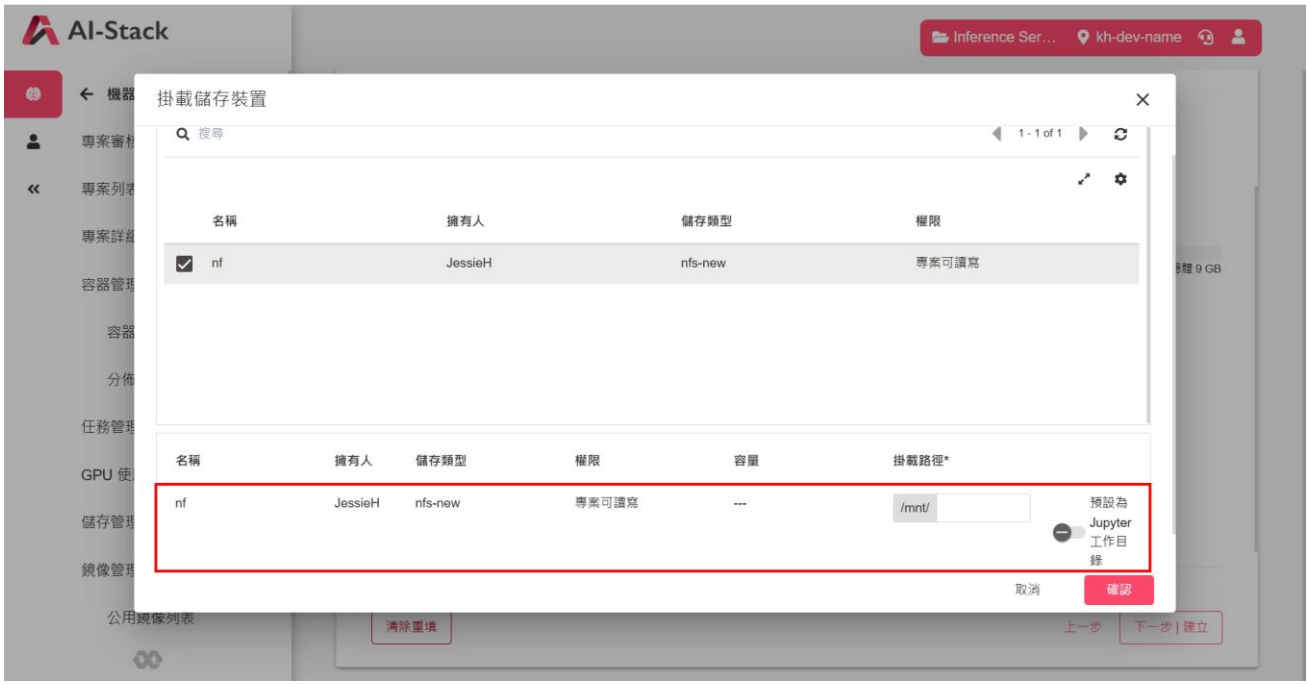
- 共享記憶體勾選是否啟用。

  共享記憶體功能說明：允許兩個或更多程序訪問同一區塊記憶體，用於多 GPU 容器進行運算時，將運算資料置於共享記憶體，提高 GPU 平行運算效率。

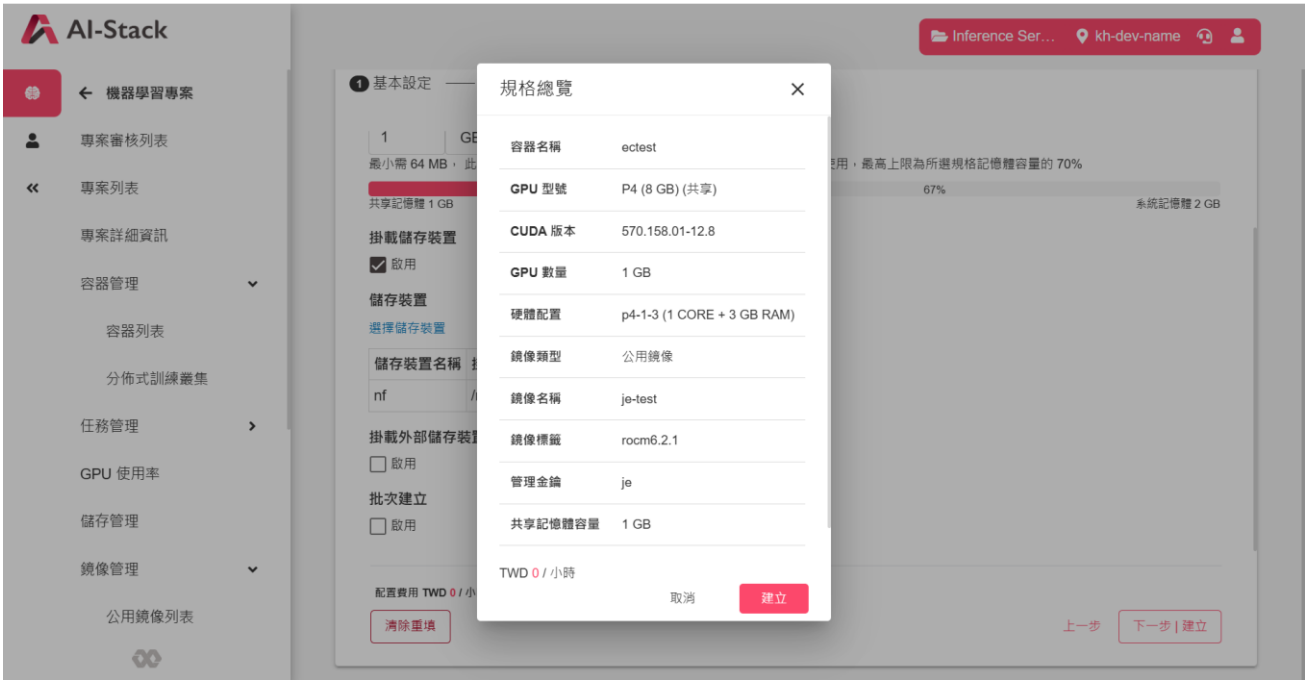
- 掛載儲存裝置。

  選取要掛載的裝置並指定掛載路徑 * ，可同時啟用 [預設為 Jupyter 工作目錄] 選項。

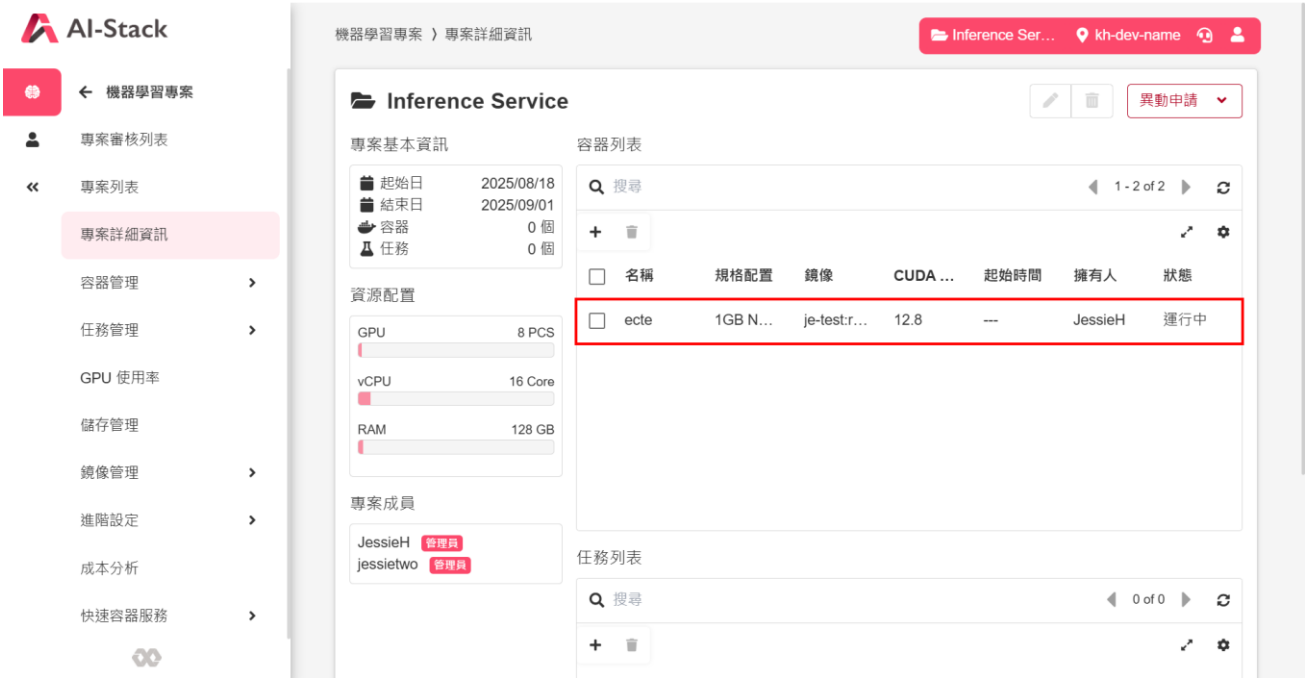
  * **備註**：需先在專案的儲存管理功能建立可用掛載裝置清單。



- 點擊 [下一步 | 建立]，即會出現規格總覽，確認無誤即可按下 [建立]。



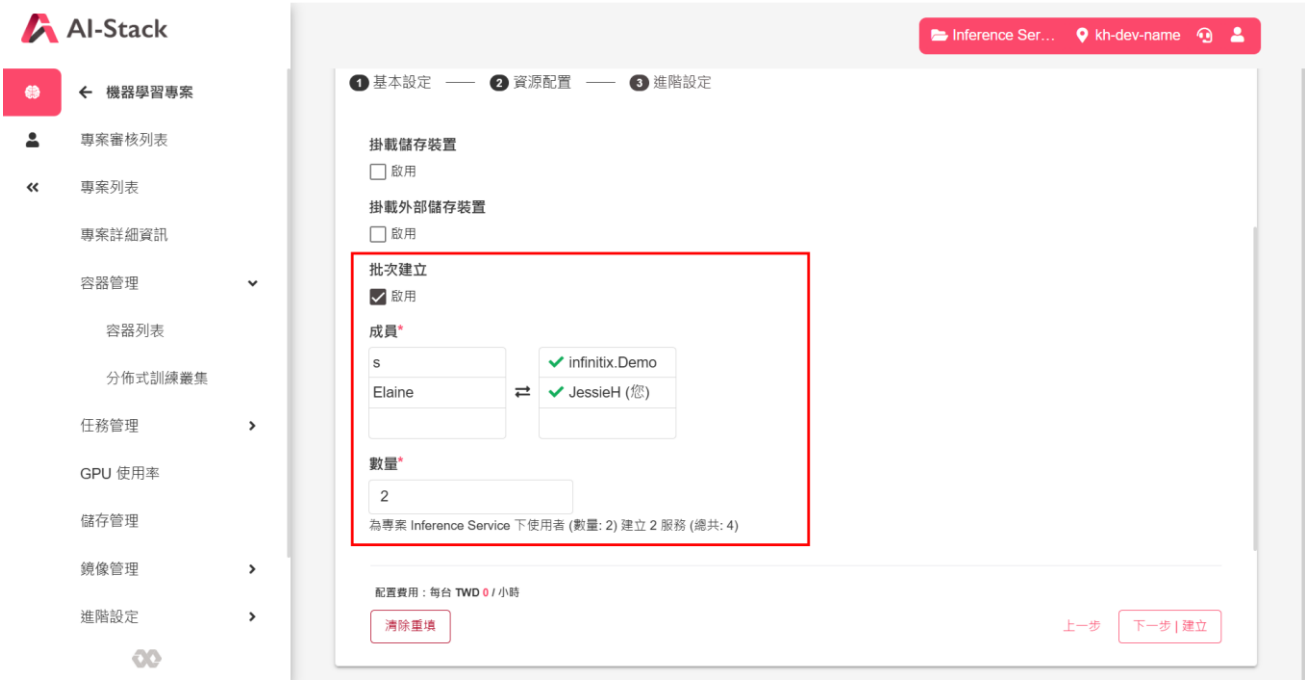
- 按下建立可以至【容器列表】及【專案詳細資訊】頁面看到建立中容器，待容器建立完成，狀態將顯示 [運行中]，即可開始使用。

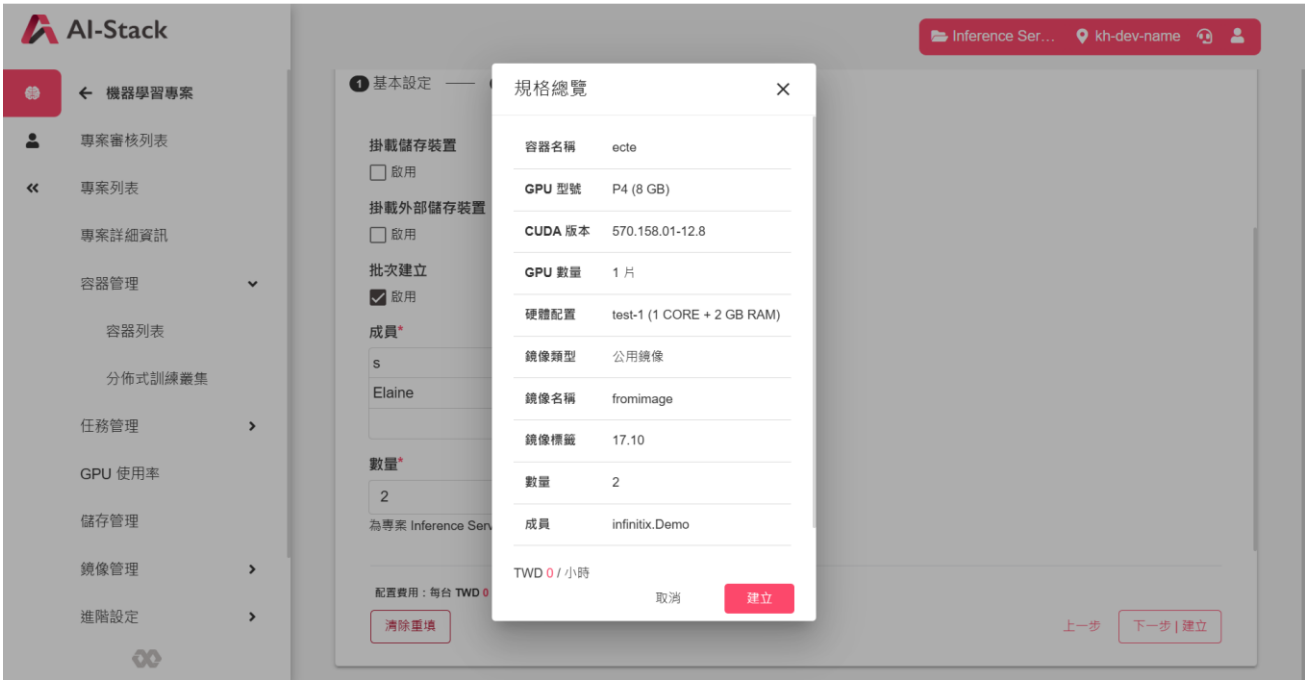


5.4.1.2 批次建立

此為專案管理者專用功能，可協助專案成員預先批量建立容器，詳細操作步驟如下：

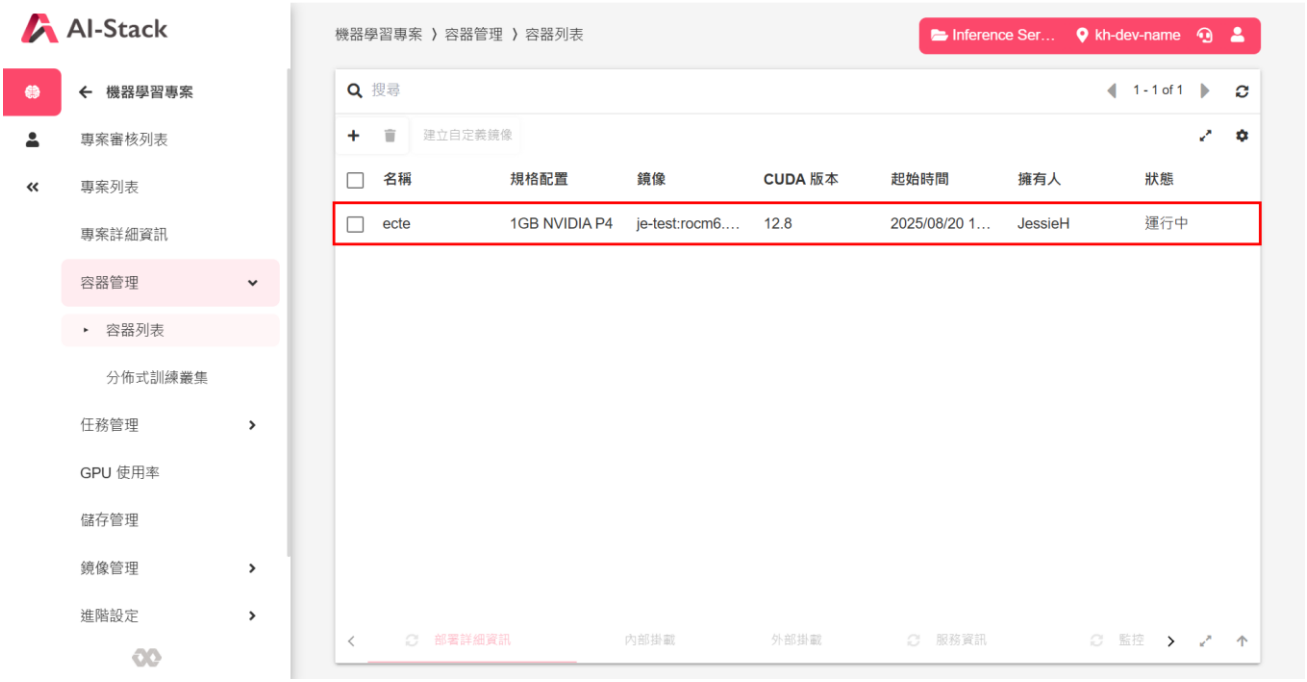
- 執行建立容器一般情形操作步驟。
- 參考下圖紅框處，在 [進階設定] 畫面中的 [批次建立] 勾選啟用。
- 將 [成員] 一欄左邊區塊人員移至右邊區塊即代表會幫他建立容器。
- 填入 [數量] 欄位，此數值代表要幫每位成員建立幾個容器。
- 點擊 [下一步 | 建立]，會出現規格總覽，確認無誤後按下 [建立]。



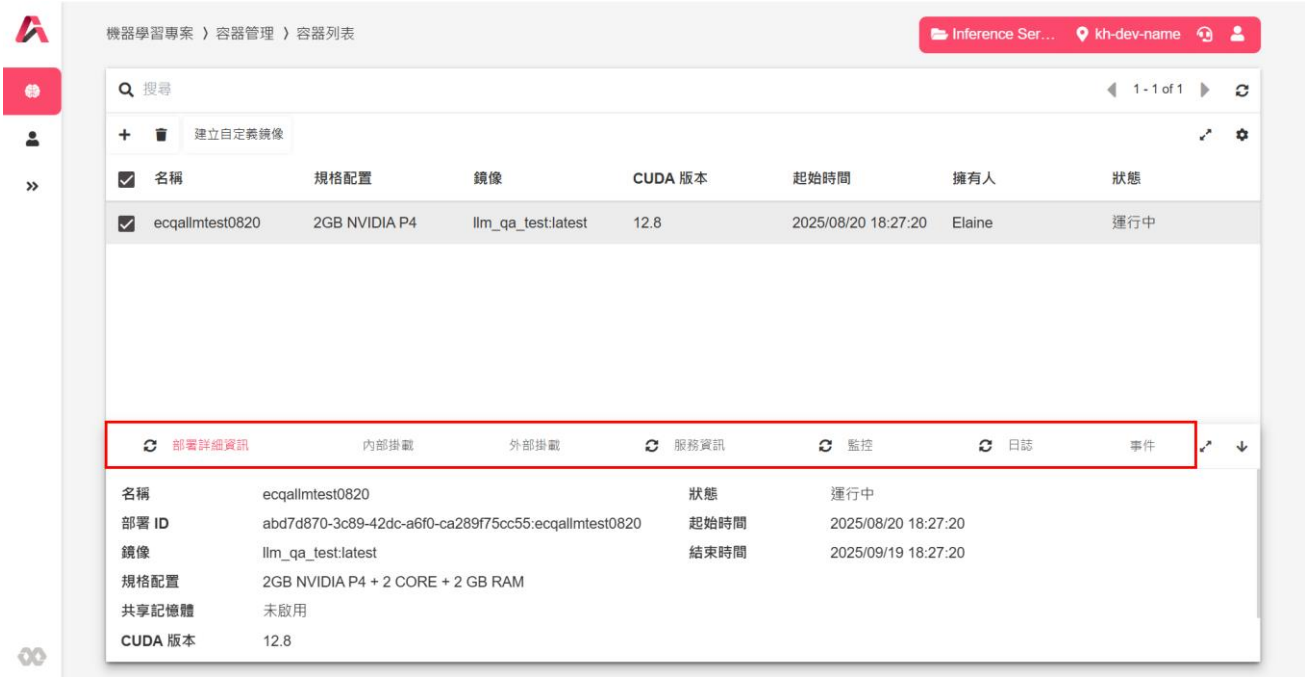


5.4.2 容器列表

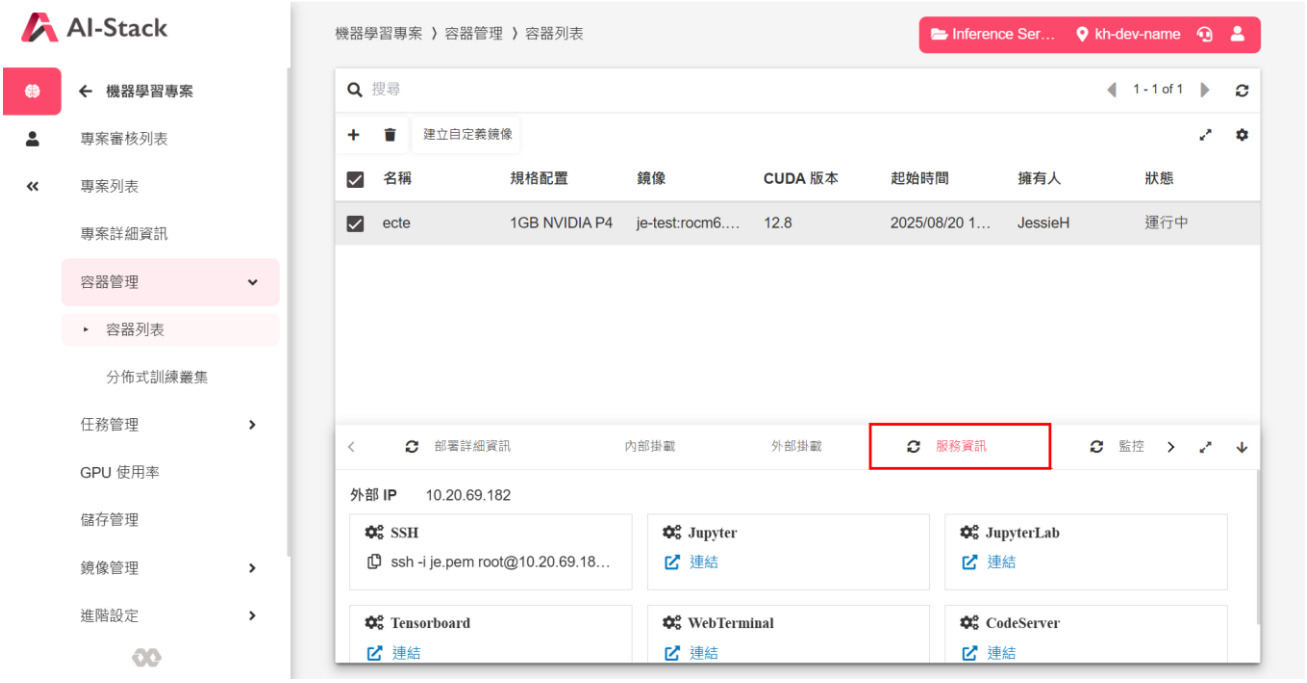
當容器建立完成後，可開啟【容器列表】查看所建立的容器清單，清單資訊包含容器名稱、規格配置（僅顯示 GPU 資訊）、鏡像、CUDA 版本、起始時間、擁有人及狀態。當運行狀態非運行中時將會以反灰顯示，無法進行任何操作。



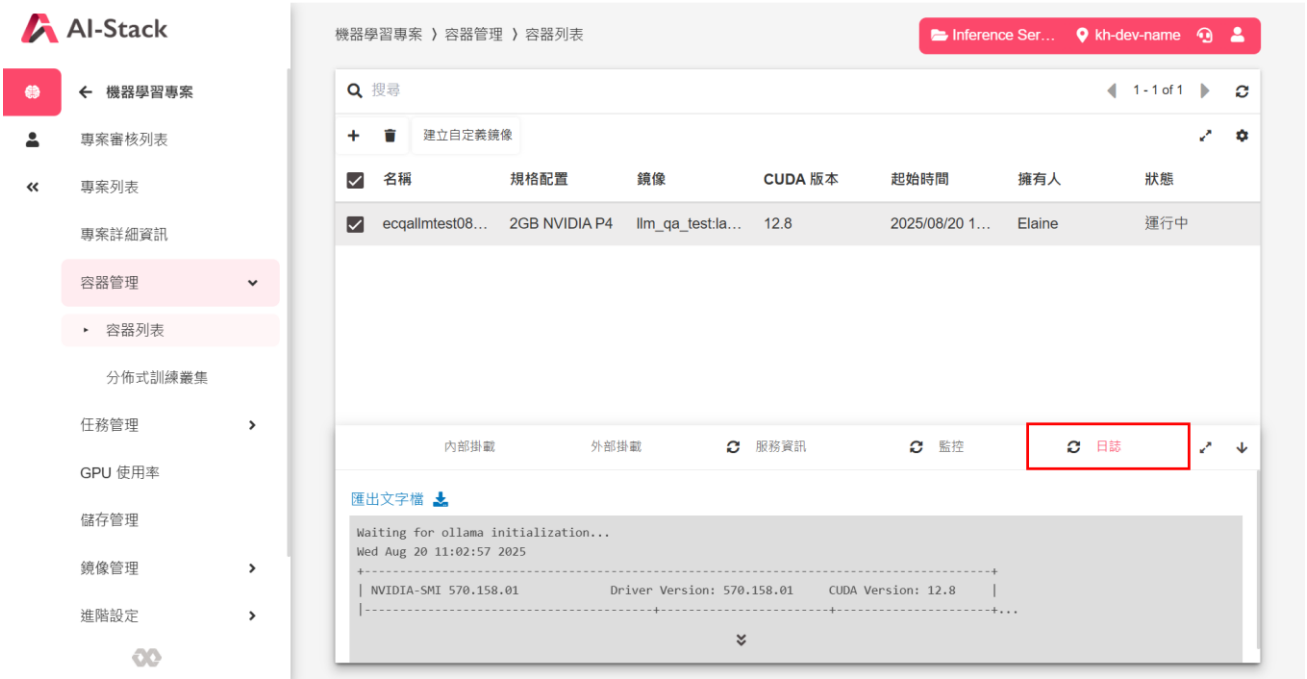
點擊狀態為運行中的容器，如下圖所示，下方將彈出包含 [部署詳細資訊] 與 [內部掛載]、[外部掛載]、[服務資訊]、[監控]、[日誌]、[事件] 等功能頁籤，提供進一步操作內容。



[服務資訊] 頁籤可查看該容器提供的所有服務，啟用方式由平台管理者在管理後台透過 MLS 樣板定義，支援 SSH、Jupyter、JupyterLab、TensorBoard、WebTerminal、Code Server 與使用者自行定義。



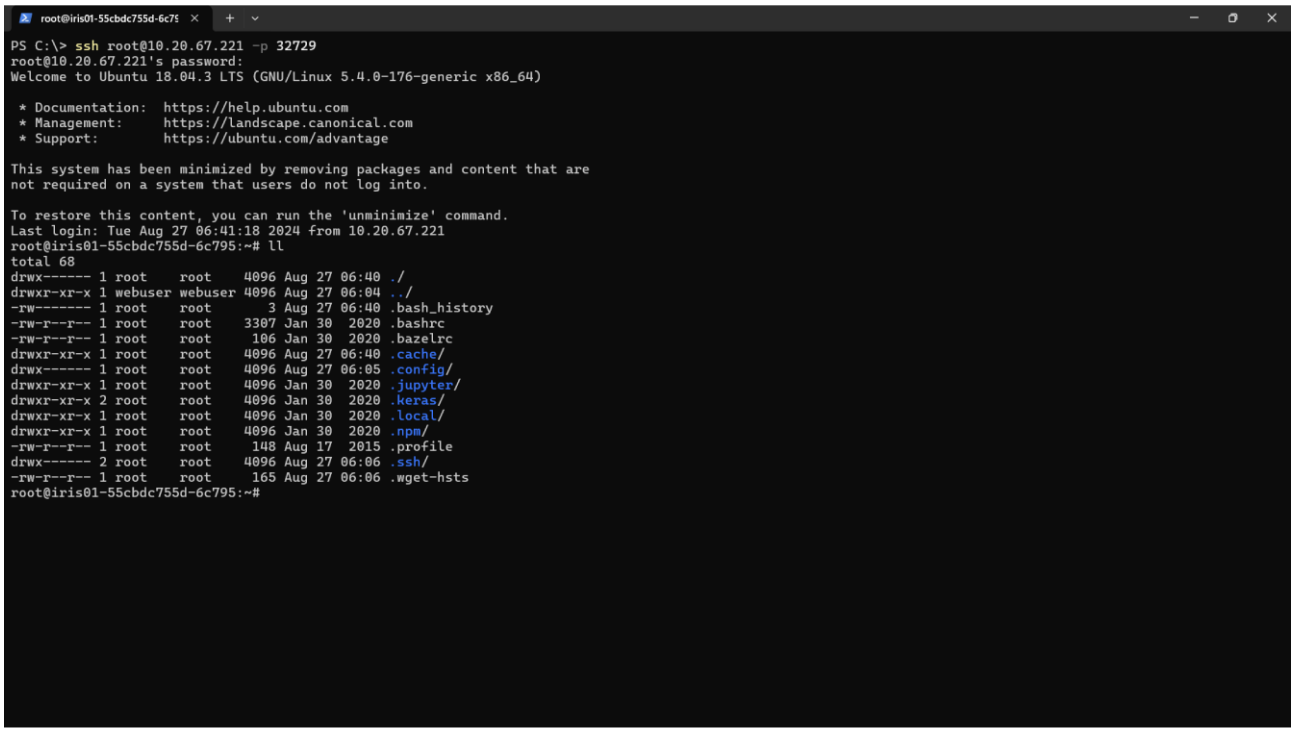
[日誌] 頁籤提供容器在運行過程中的所有日誌輸出，包括系統訊息、應用程序運行紀錄以及錯誤報告等詳細內容。用戶可以透過查看這些日誌，瞭解容器的運行情況、執行過程中的關鍵步驟以及可能發生的異常情況。



5.4.2.1 容器服務使用方式

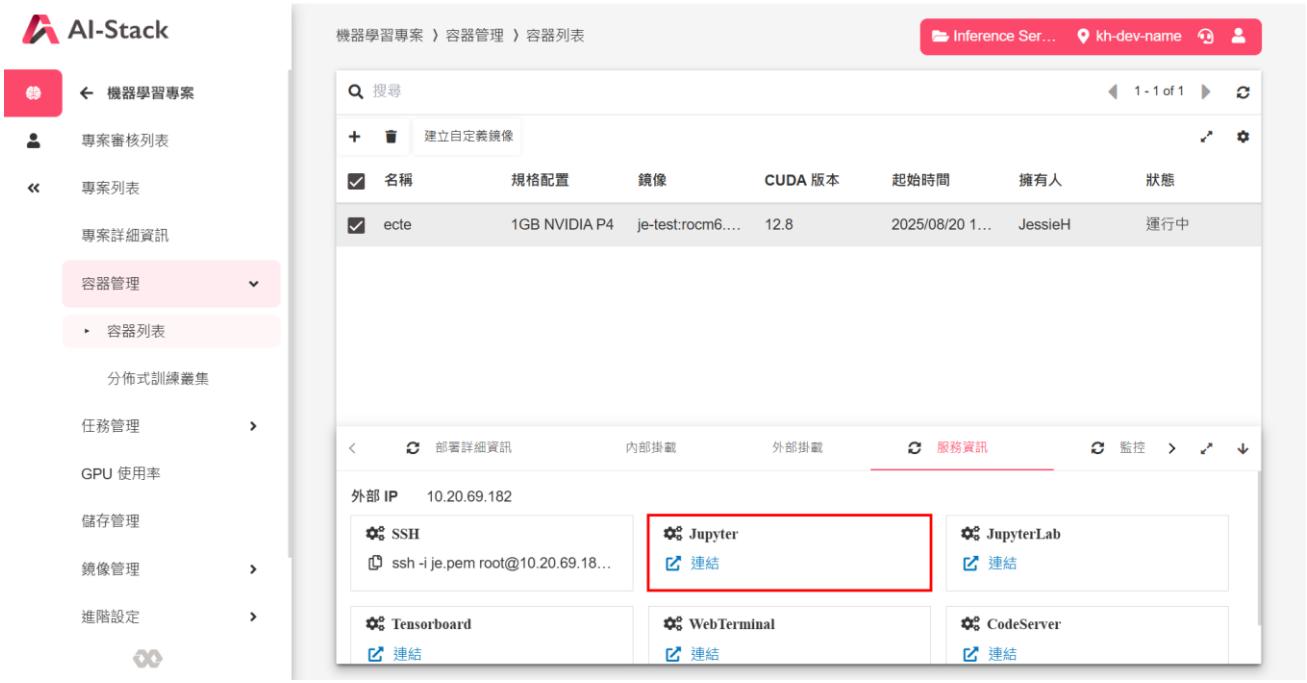
- SSH

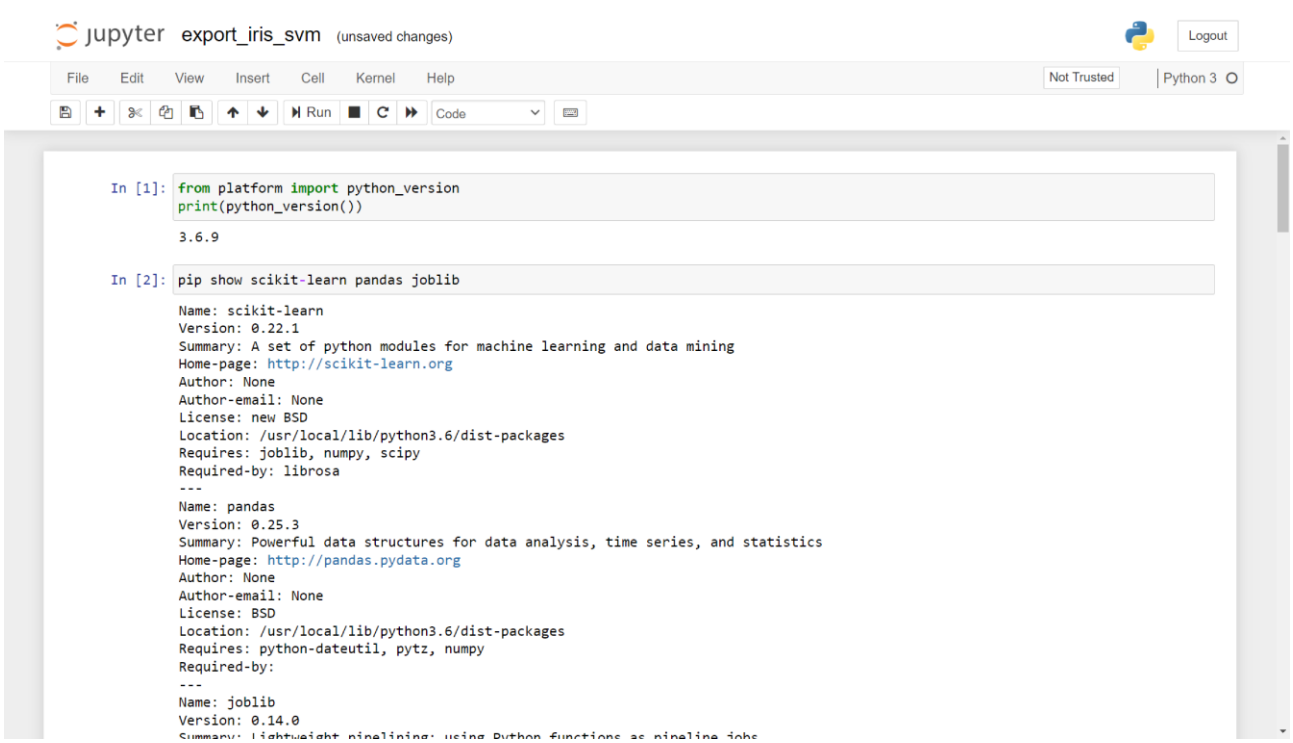
  使用 SSH 工具（如 PuTTy）配合 IP、服務埠號（port）及建立時所輸入的密碼或選擇的金鑰，即可以 SSH 遠端登入容器內進行操作。以查看前述開發中容器詳細資訊為例，SSH 指令為「ssh root@10.20.67.221 -p 32729」。



- Jupyter

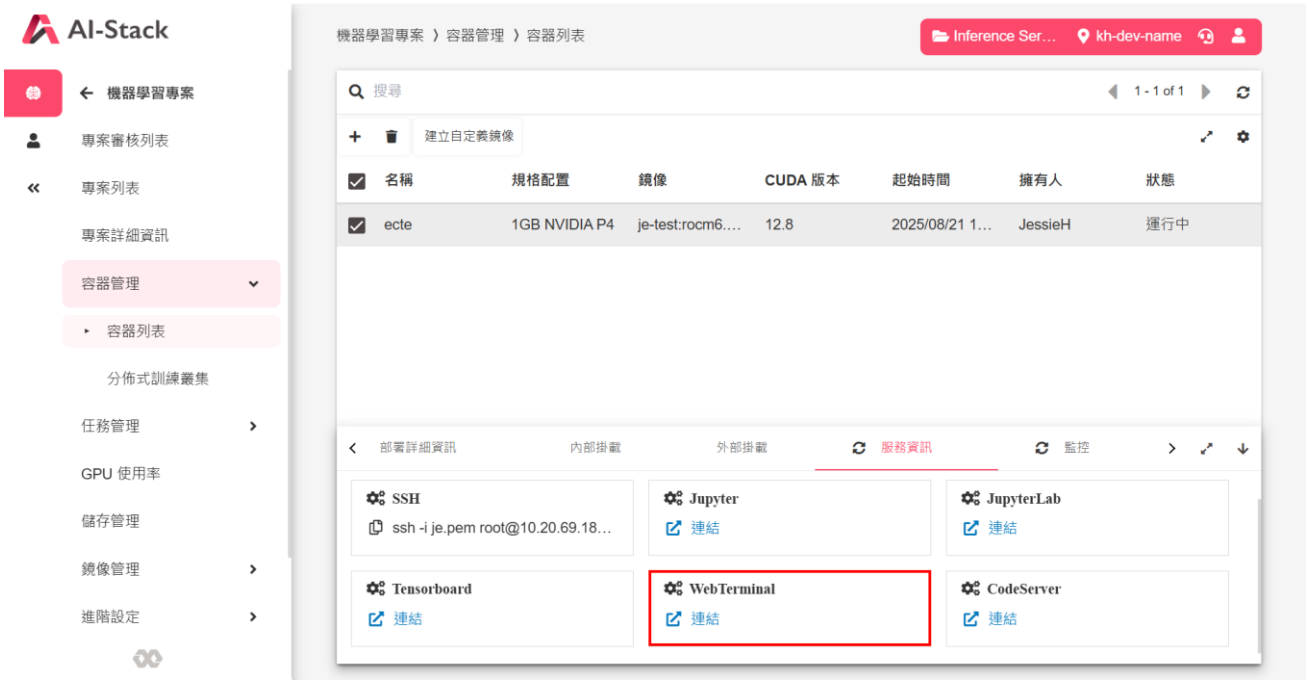
  於 [服務資訊] 頁籤中，點擊 Jupyter 連結按鈕將直接開啟 Jupyter Notebook，使用者可透過其編寫 Python 程式。

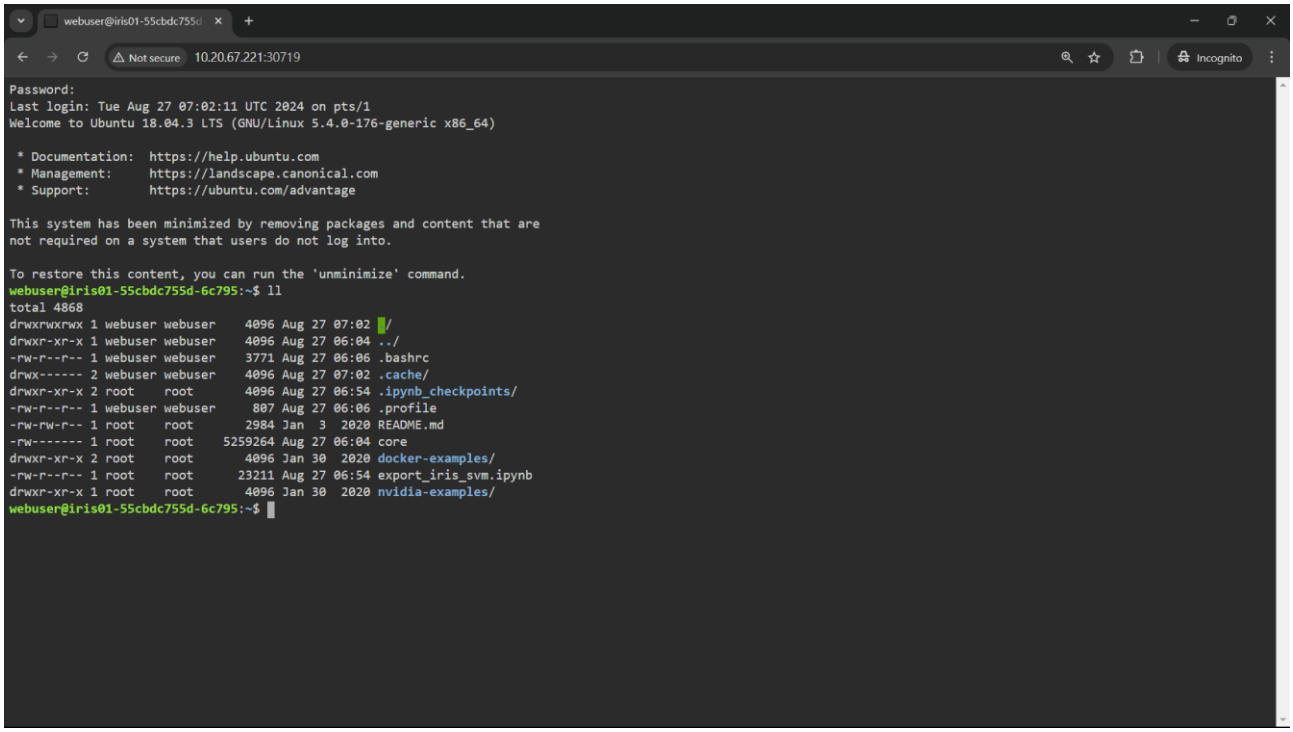




- WebTerminal

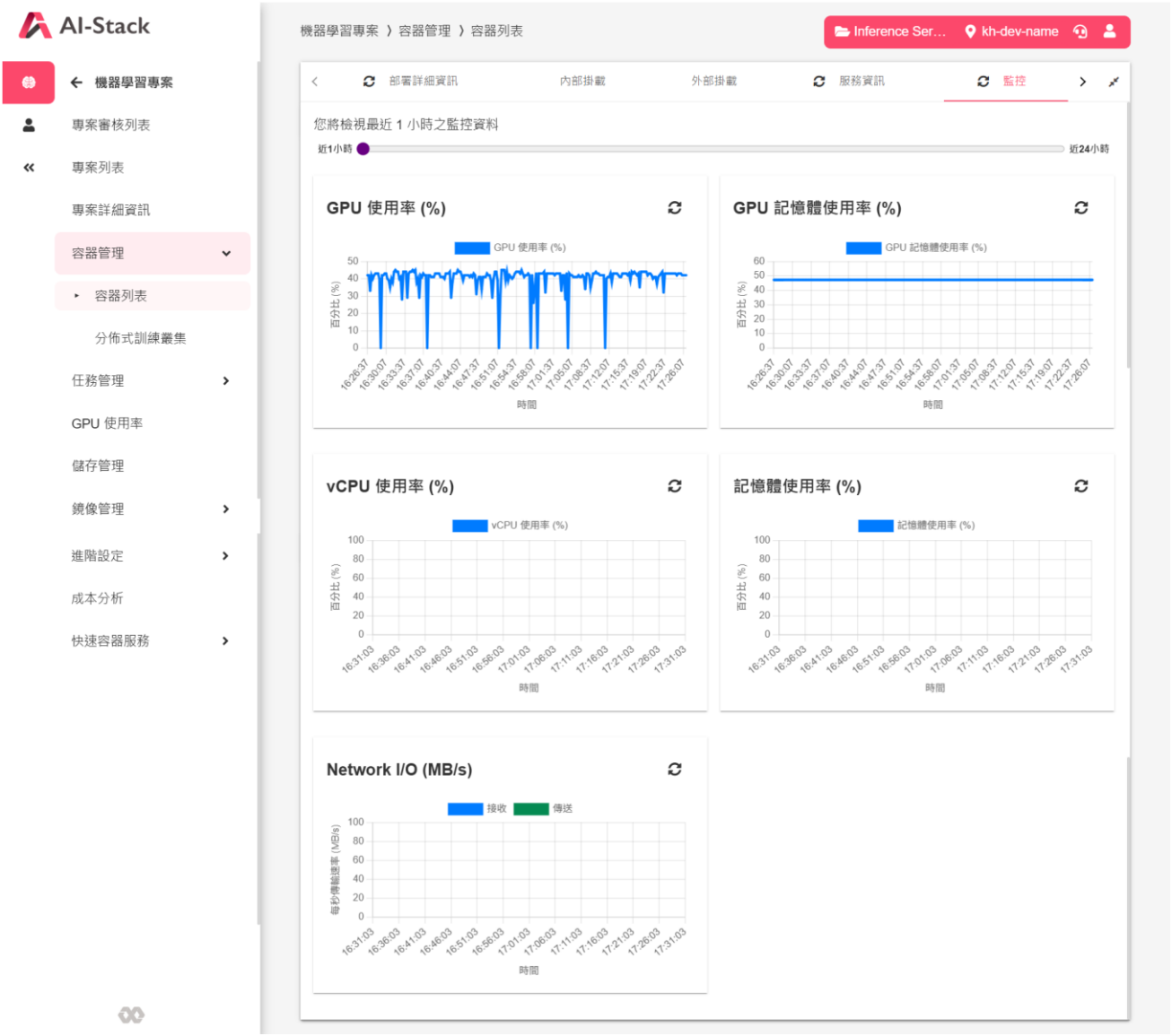
  於 [服務資訊] 頁籤中，點擊 WebTerminal 連結按鈕將直接開啟 WebTerminal，可直接進行容器系統環境操作，如下圖。





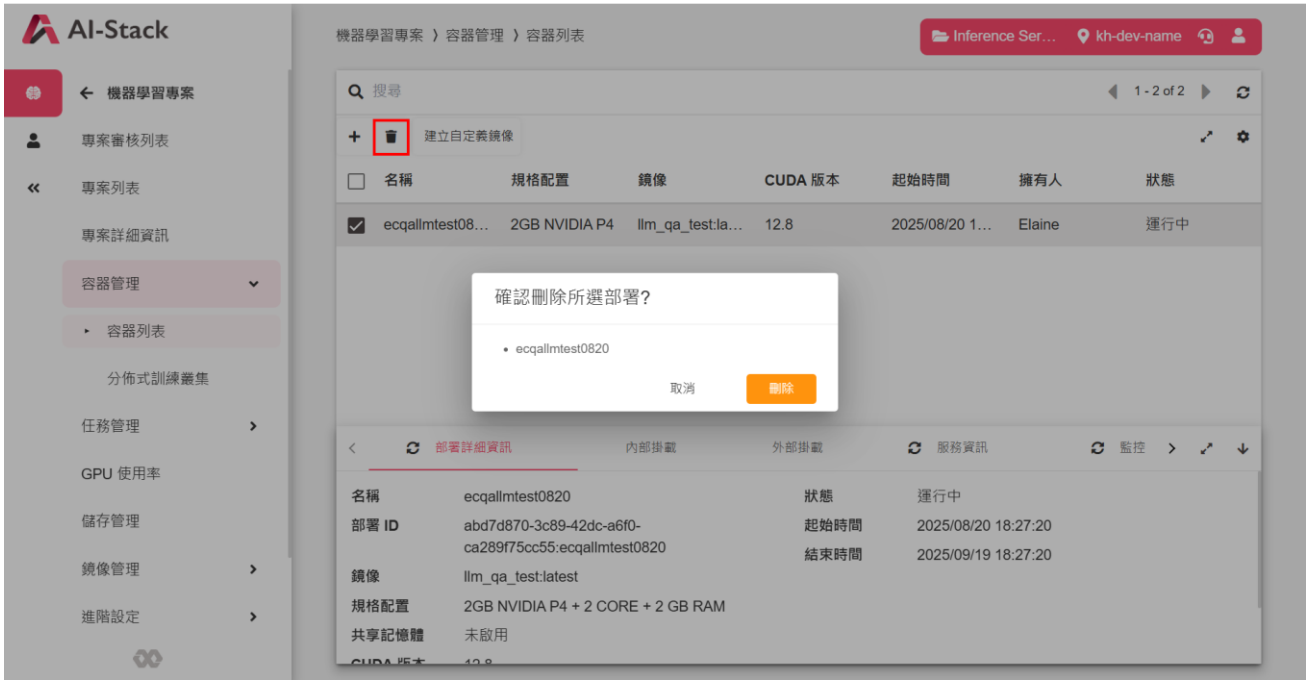
5.4.2.2 監控

使用者能透過 [監控] 頁籤來查看容器之 GPU 使用率、GPU 記憶體使用率、CPU 及記憶體使用率、Network I/O 等圖表，也可以透過拖動上方時間軸 (累積使用時間超過 1 小時才會顯示) 切換要檢視 1 至 24 小時內，哪一種區間的數據呈現。



5.4.3 刪除容器

欲刪除容器時，可於清單中勾選目標容器，選定後點擊  將出現確認畫面，如下圖所示，確認為想要刪除的容器後再點擊 [刪除]。



* 注意：容器一旦刪除即無法再行恢復，請務必確認已備份所需檔案及資料，或將需要保留之資料另行存放至掛載裝置路徑底下，確保資料可再取回。

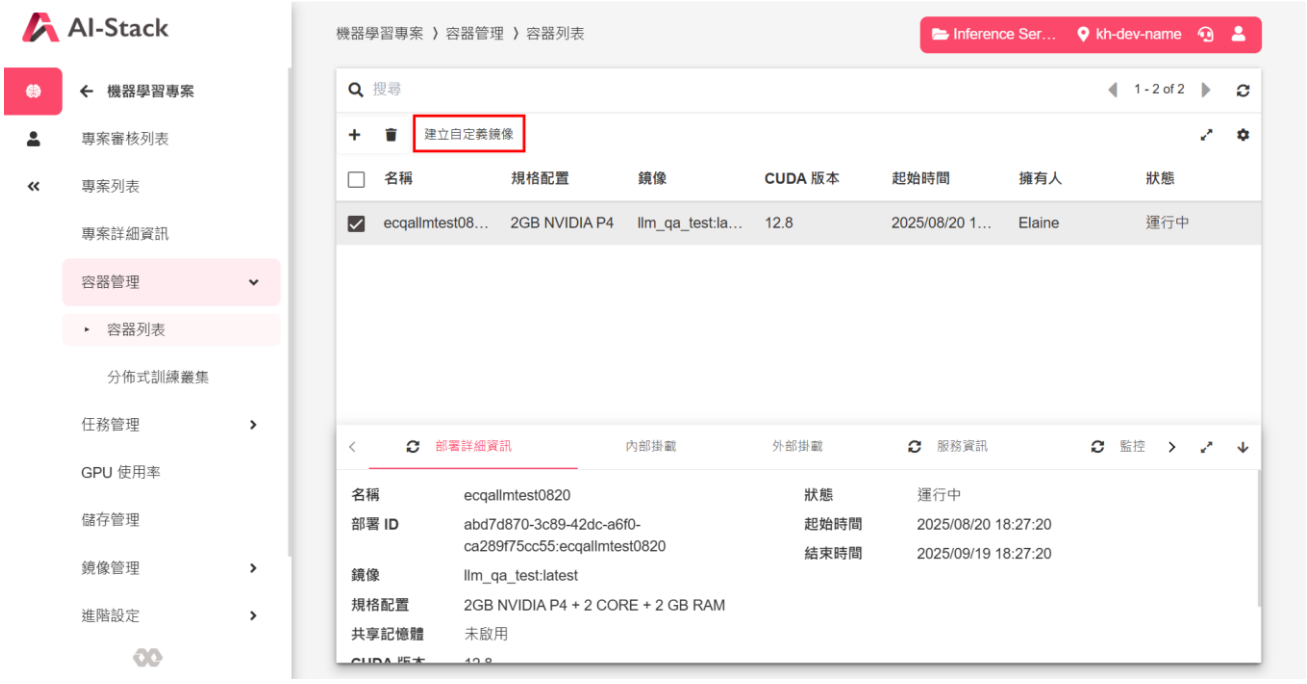
5.4.4 建立自定義鏡像

建立自定義鏡像目的為提供使用者可自行容器樣板，例如取用管理者上架的公用鏡像後，再依個人需求，安裝額外套件，並製作成自定義鏡像，方便快速還原個人化開發環境設定。操作方式為在【容器列表】中，選取目標容器後，再點選列表上方 [建立自定義鏡像] 按鈕，即可帶出設定功能頁籤。

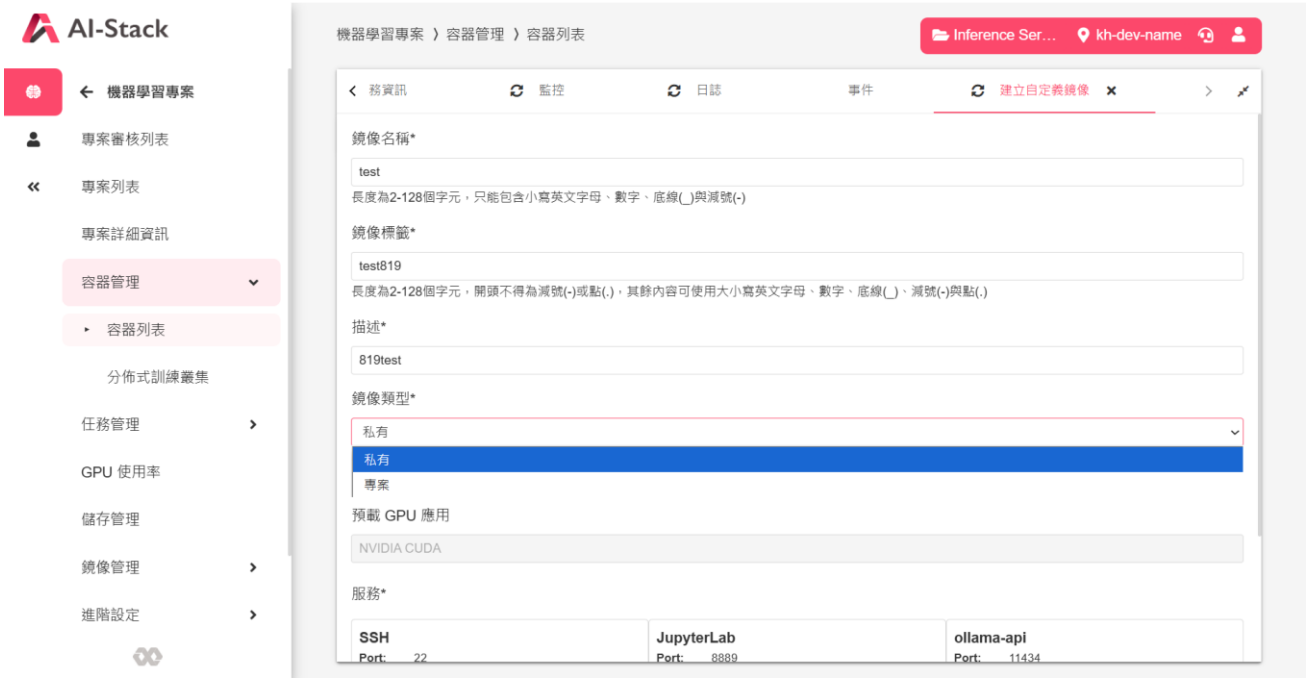
使用者可填入易於識別的鏡像名稱、鏡像標籤及描述，以建立所需要的鏡像環境，也可以透過鏡像類型之選項，指定建立後的鏡像為個人使用 ([私有])，或是為專案內的所有成員共同使用 ([專案])。

5.4.4.1 建立方法

- 選擇要轉為自定義鏡像的容器環境，然後點擊 [建立自定義鏡像]，如下圖。



- 填入鏡像名稱、鏡像標籤（這兩個欄位將同步用於鏡像儲存庫）。
- 填入描述。
- 選擇鏡像類型，該選項可以指定建立後的鏡像為個人使用（私有），或是為專案內的所有成員共同使用（專案）。

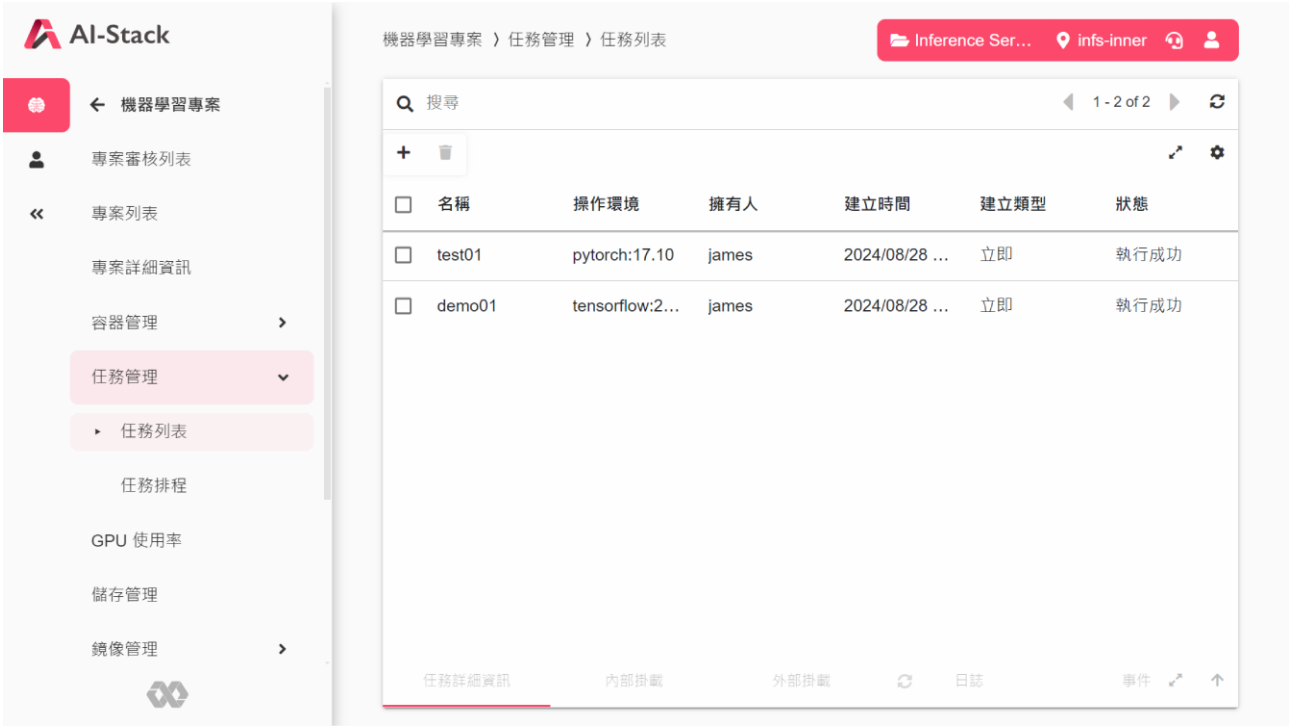


建立完成的鏡像將出現於【鏡像管理】>【自定義鏡像列表】列表底下，如下圖。

5.5 任務管理

5.5.1 任務列表

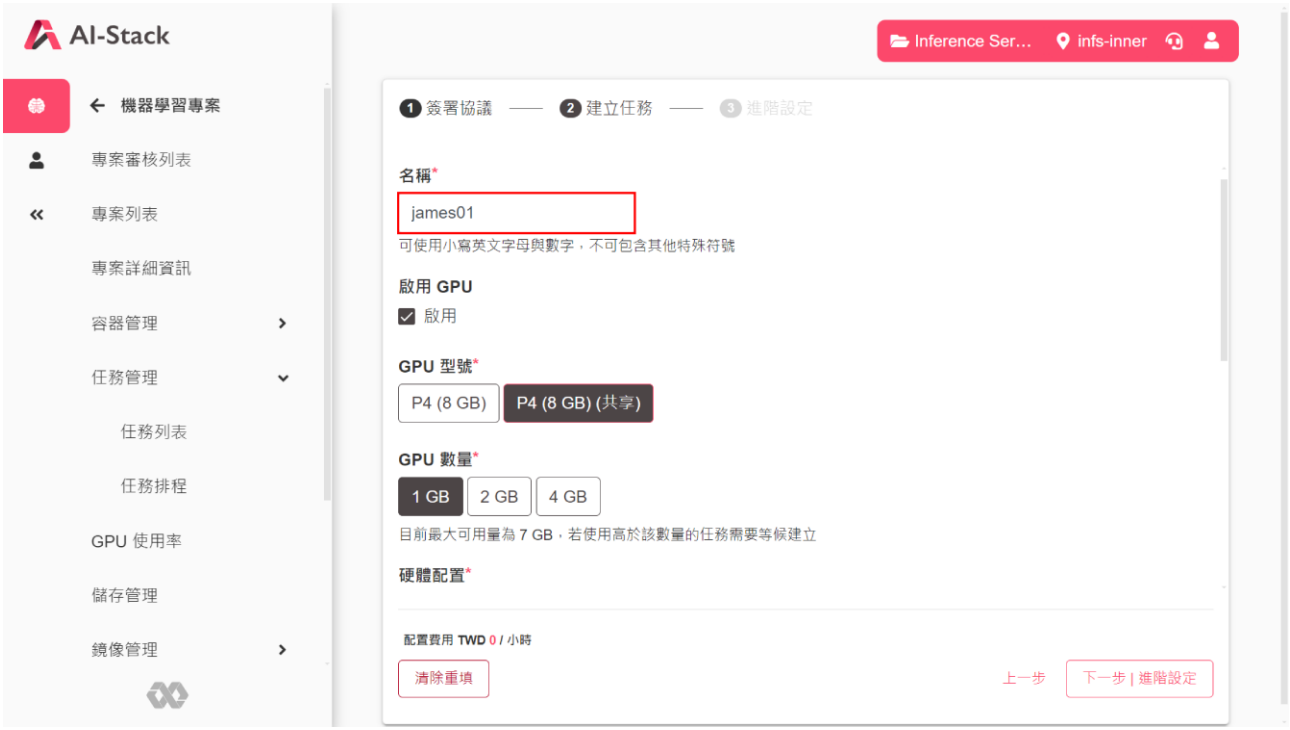
於左側選單點選【任務管理】>【任務列表】即可查看任務列表清單：



- 建立任務

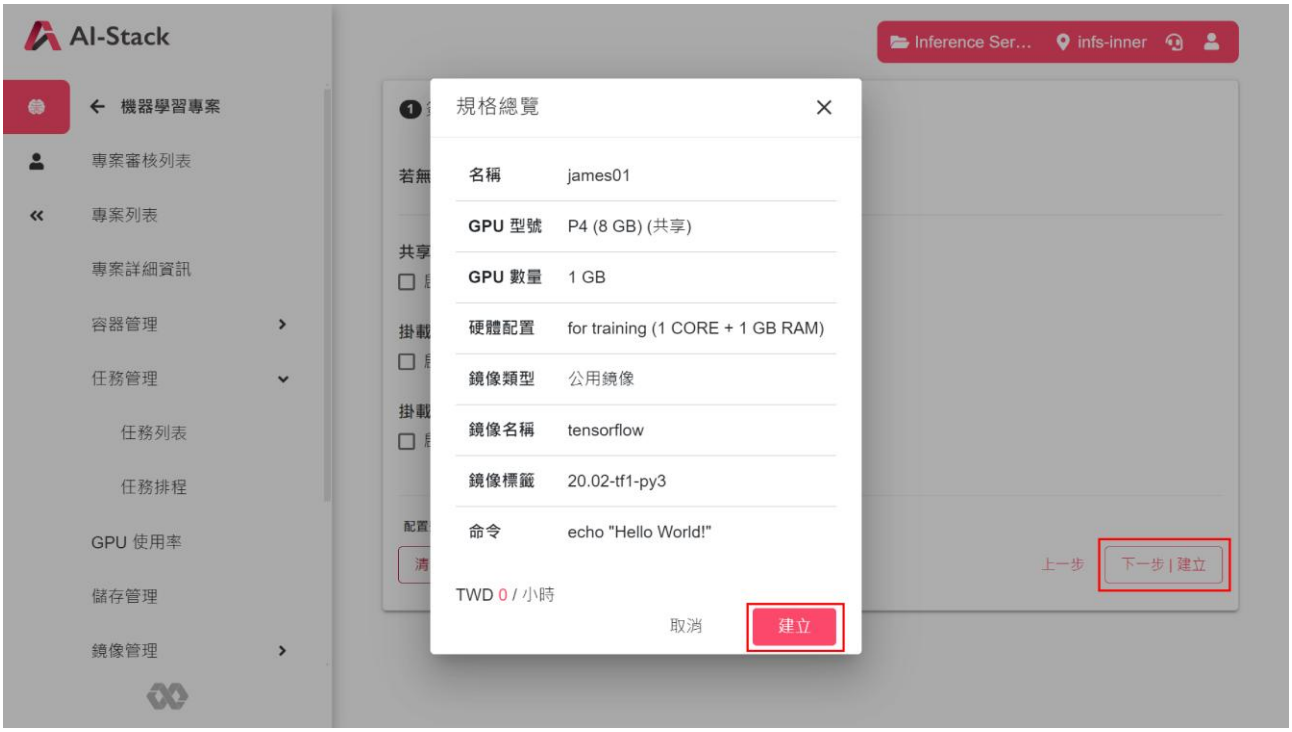
  任務的建立方式如下：

  - 開啟任務列表畫面，點擊列表左上方 ⊞ 圖示。
  - 在建立任務畫面輸入名稱。
  - 選擇任務容器 GPU 型號、數量、硬體配置 (含 CPU 核心數、記憶體)。

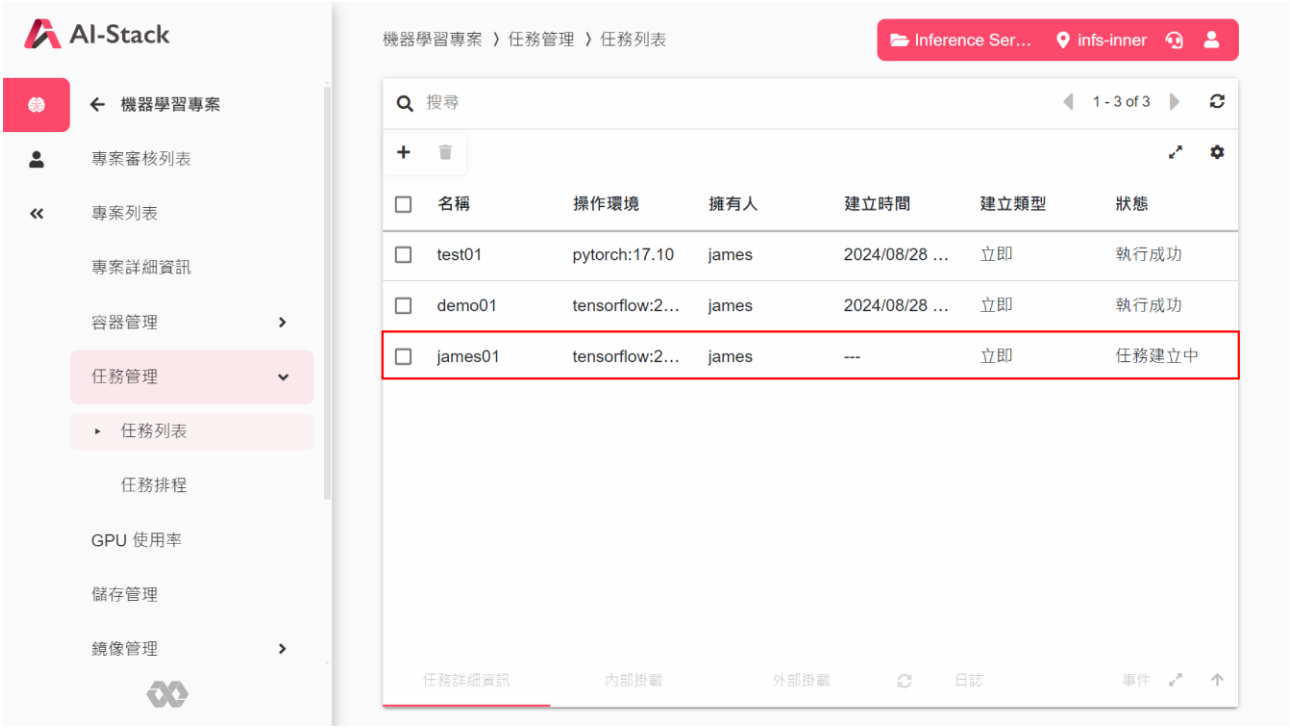


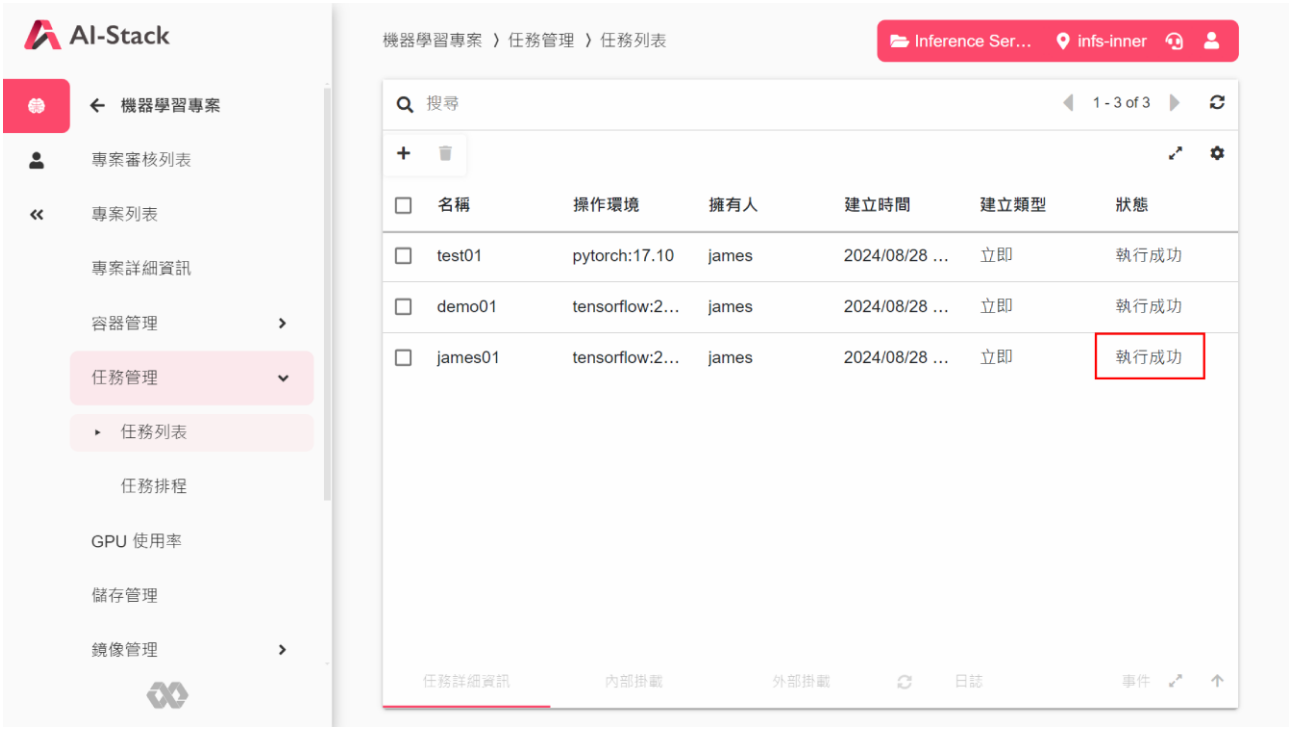
- 選擇鏡像。
- 命令一欄請輸入該任務欲執行之命令。
- 點擊 [下一步 | 進階設定]。

- 若不需啟用進階設定的功能，可直接點擊 [下一步 | 建立] 跳出規格總覽頁面。
- 確認無誤後點擊 [建立] 即可送出。



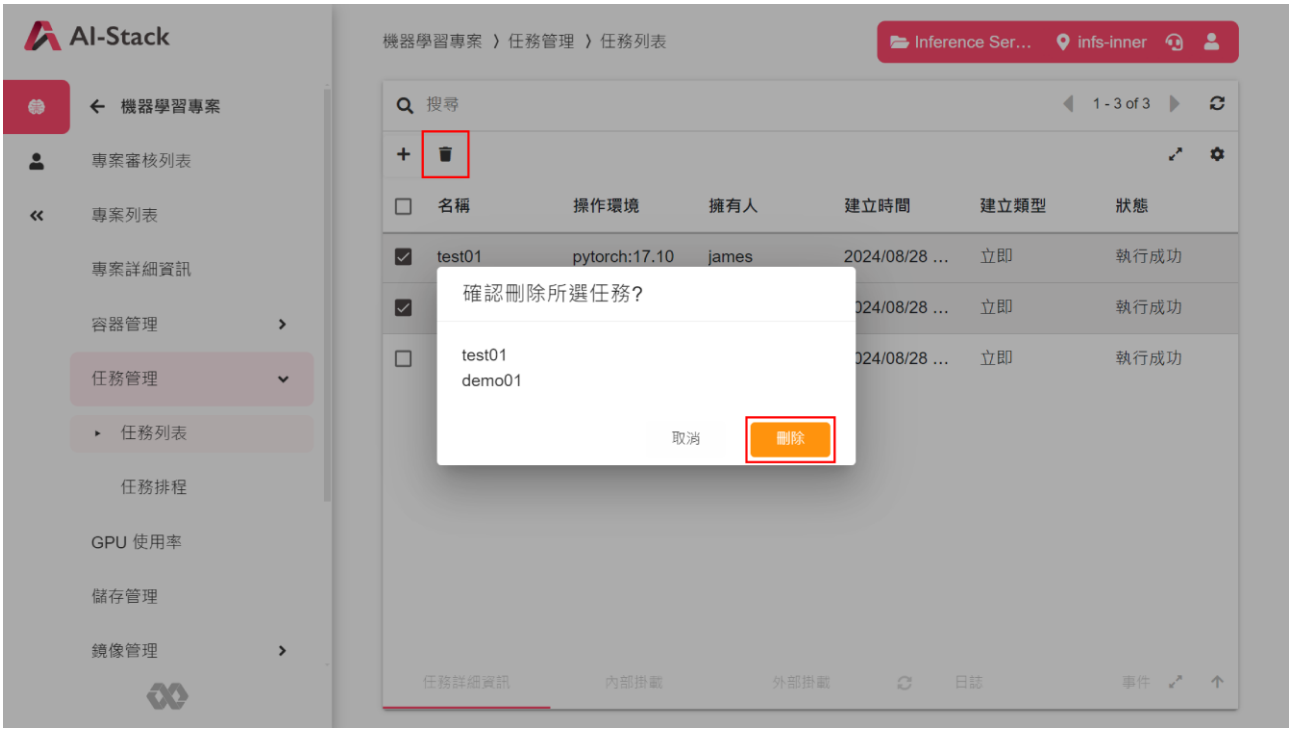
■ 建立過程將呈現在任務列表狀態一欄，直到更新為 [執行成功]。





- 刪除任務

已經不需要的任務，可透過任務列表畫面勾選（可多選），再點選  圖示進行刪除。



5.5.2 任務排程

以下兩種方式可以設置任務排程：

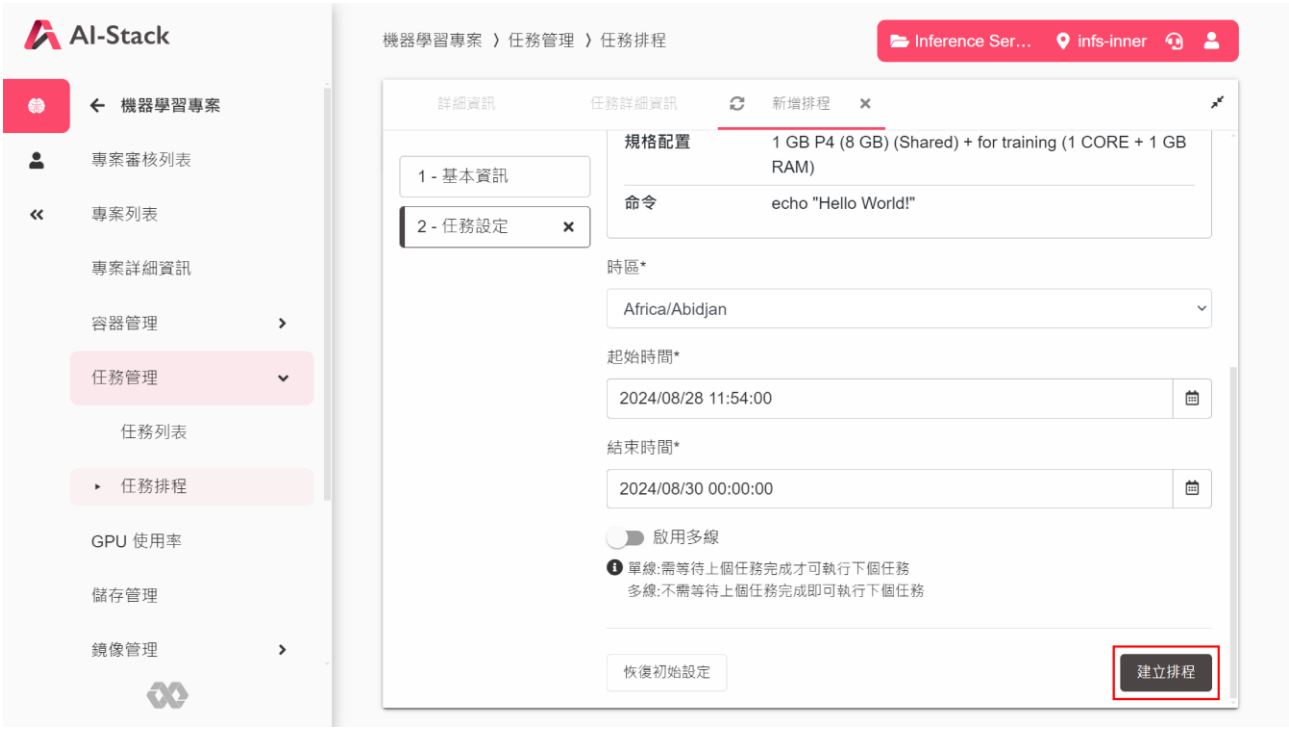
1. 從【任務管理】>【任務排程】頁面，先設定排程，再關聯任務樣板。

- 於左側選單選擇【任務管理】>【任務排程】。
- 點擊左上 ⊞，開啟新增排程頁面。
- 輸入排程名稱、描述。
- 選擇觸發時間 (即觸發頻率)，如下圖示意為每月、每天 20：10 執行。
- 點擊 [添加任務至排程]，帶出任務設定畫面。

- 任務目標下拉選擇已先建立好的 [任務樣板] * ，下方將帶入規格詳細資訊。

  * 任務目標選項需先於【進階設定】>【任務樣板】進行設定，操作方式請參考建立任務樣板。

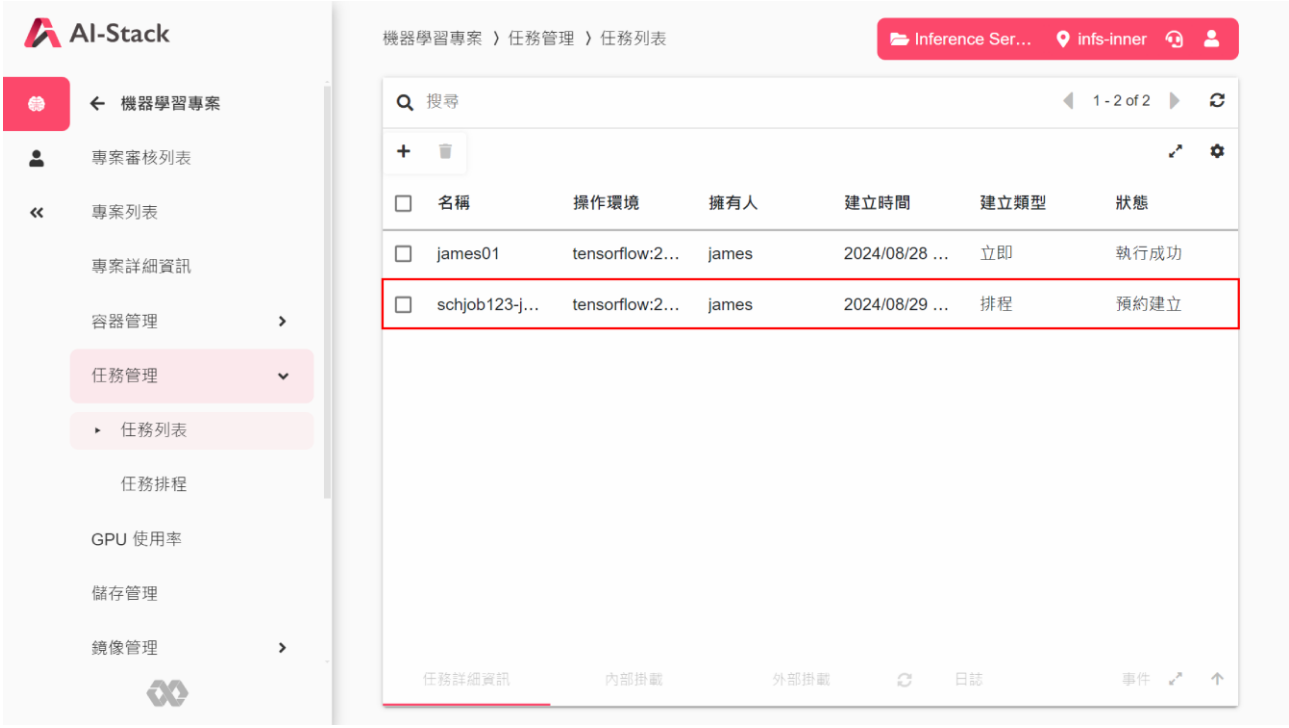
- 選擇時區、起始時間及結束時間。
- 設定是否 [啟用多線]。
  - 單線：每次執行需等待上一次啟動的任務完成，才會啟動下個任務。
  - 多線：此排程任務無需等待，將直接啟動。
- 確認填寫內容後，點擊右邊 [建立排程] 按鈕。



- 建立成功將出現在任務排程列表。

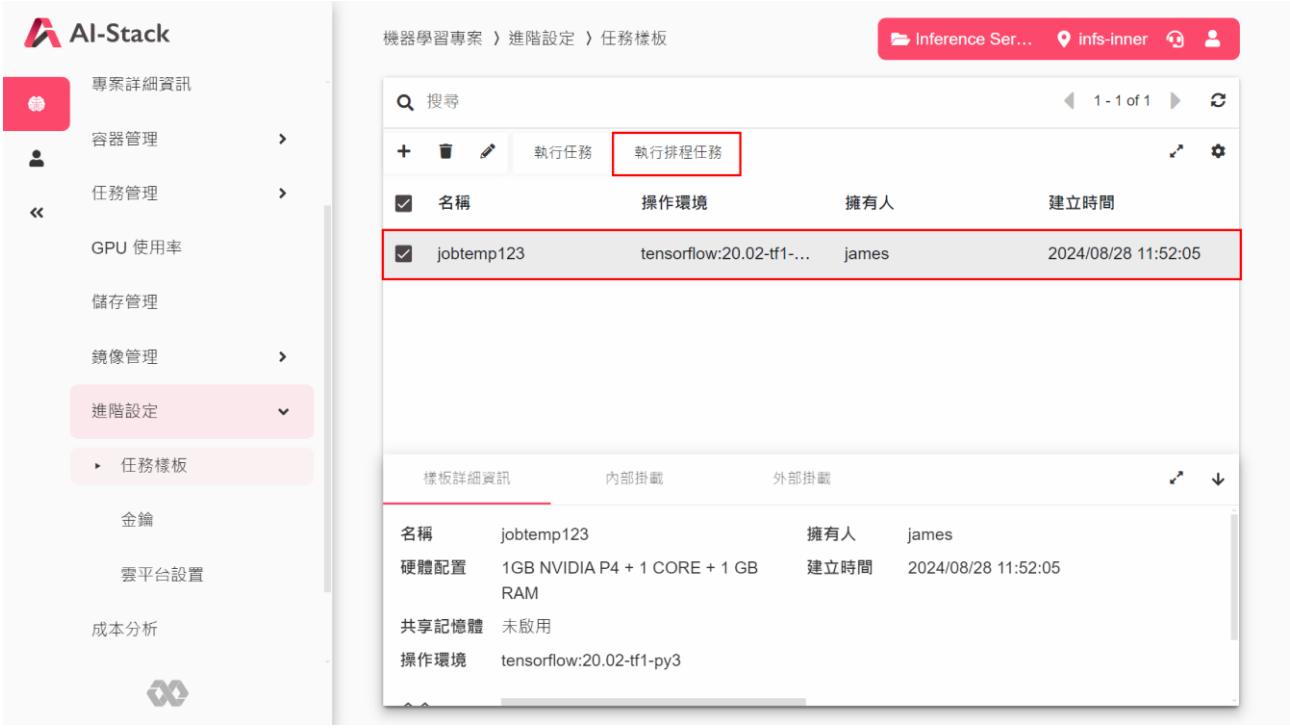
- 任務列表中亦將新增一筆即將觸發的任務 * 。

  * 注意：該筆任務將無法刪除，若要刪除必需將該任務對應的任務排程整筆進行刪除，才可一併刪除此任務。

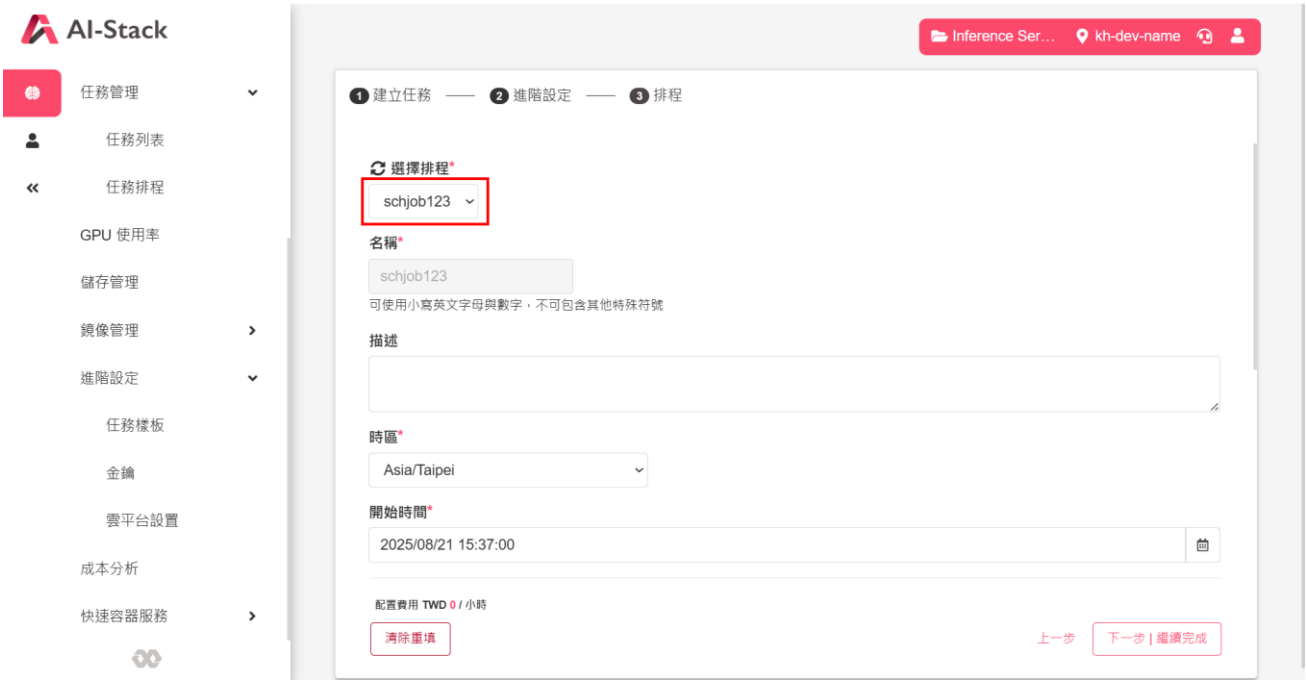


2. 從【進階設定】>【任務樣板】頁面，先選擇任務樣板，再關聯或建立排程。

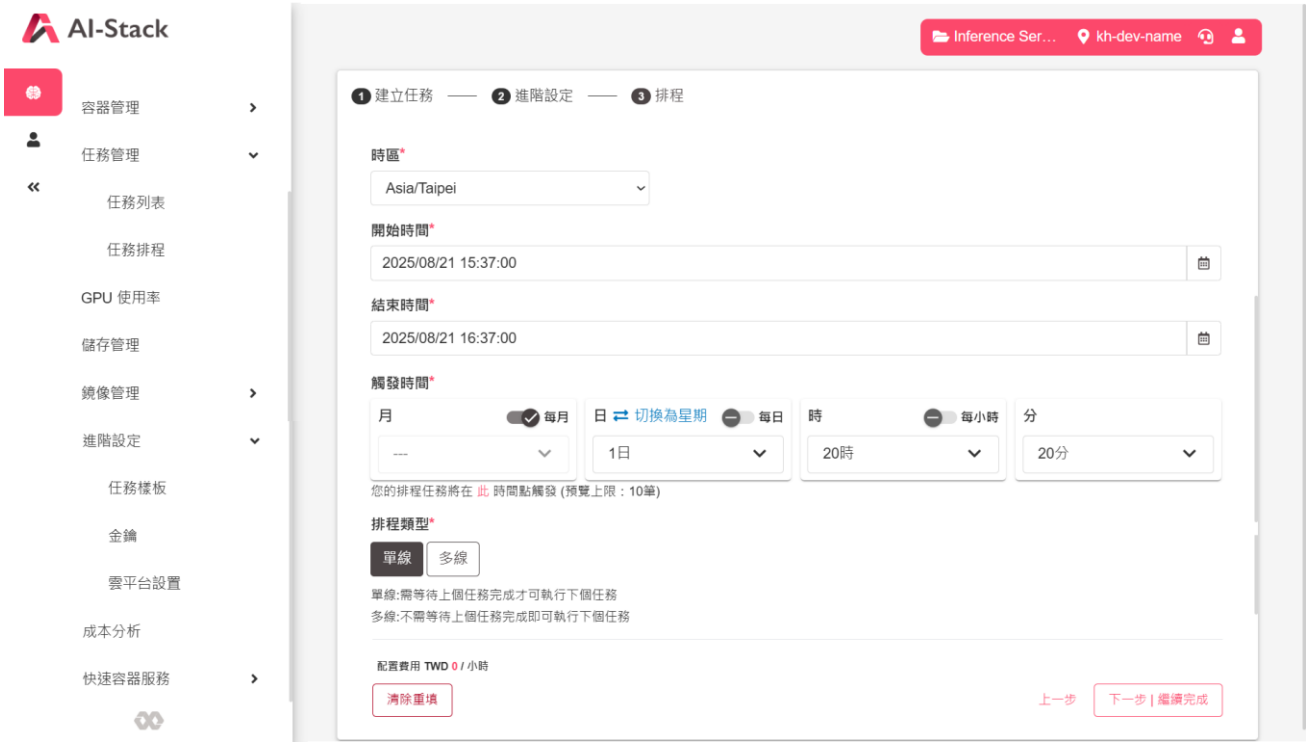
- 於左側選單開啟【進階設定】>【任務樣板】。
- 勾選列表中指定的任務樣板，點擊 [執行排程任務] 按鈕。



- 進入設定排程頁面，可選擇新建排程或開啟既有排程進行編輯。
- 如選擇既有的排程設定，系統將帶出排程設定資料，名稱將不能更改。

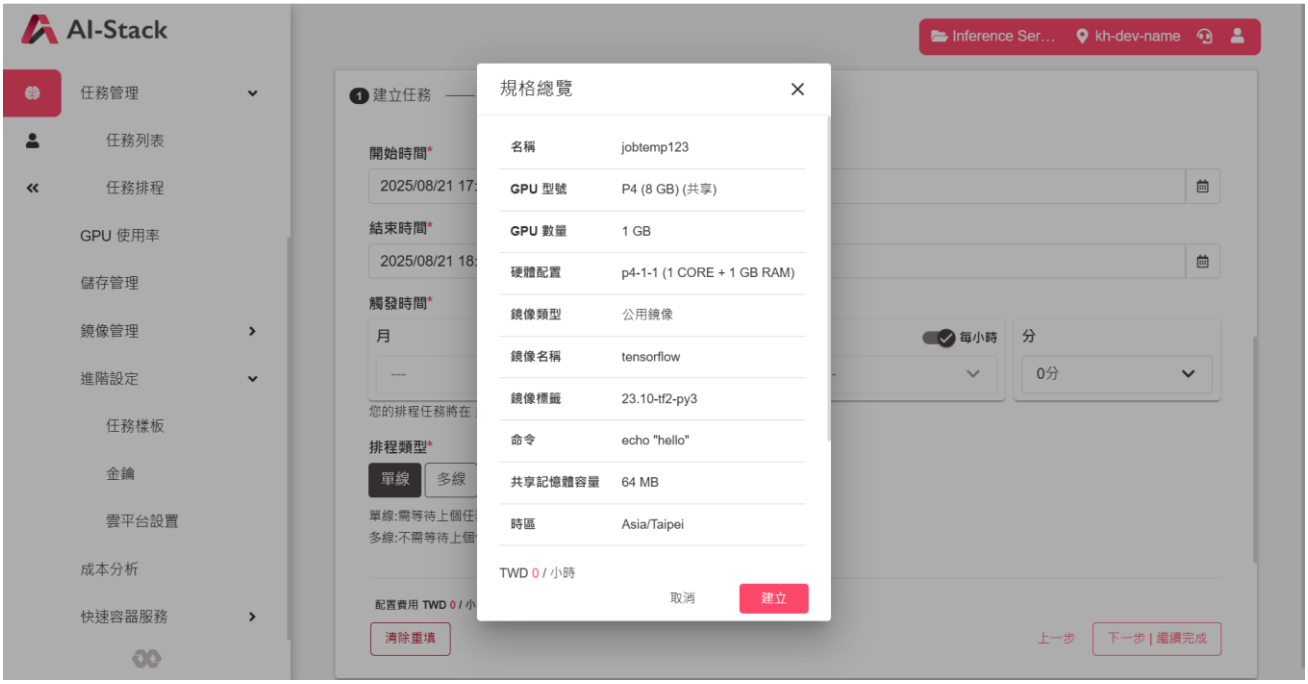


- 確認時區、開始時間、結束時間、觸發時間、排程類型選項後，點擊 [下一步 | 繼續完成]。

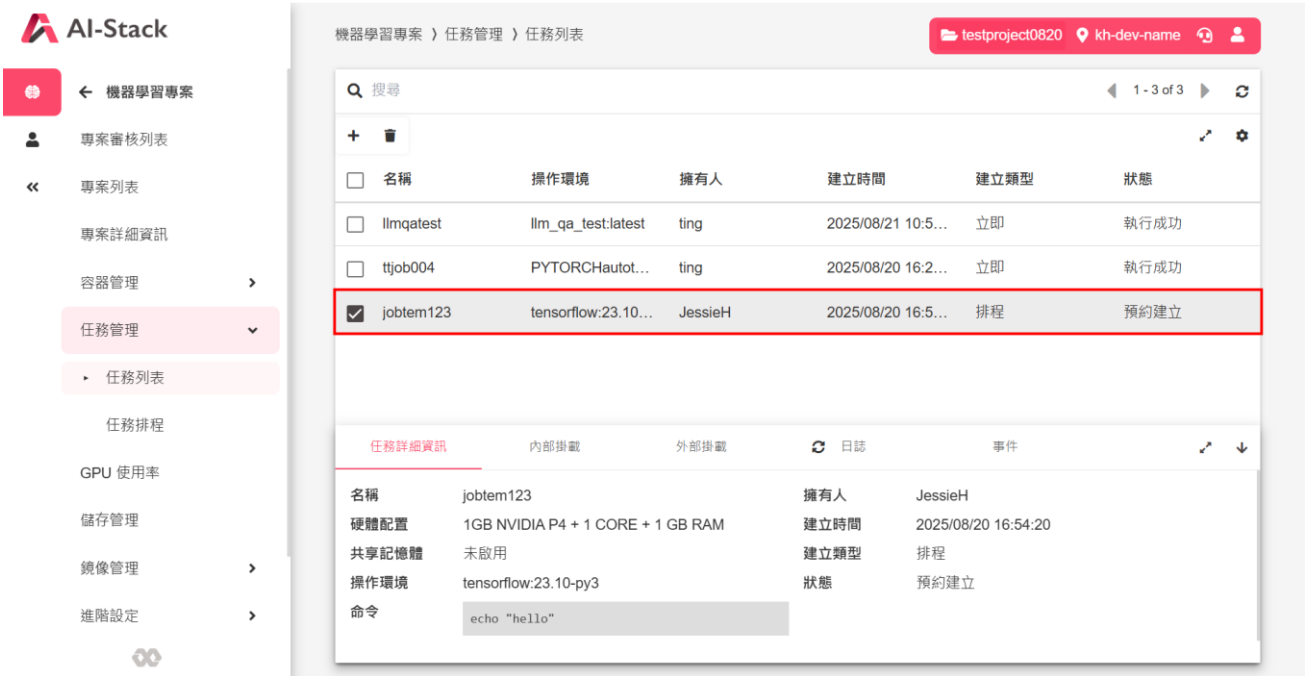


- 如選擇建立新的排程設定，則需從頭填入名稱、開始時間、結束時間、觸發時間、排程類型選項，填寫完成後點擊 [下一步 | 繼續完成]。

- 跳出的規格總覽視窗，如確定填寫資料沒有問題，點擊 [建立]。

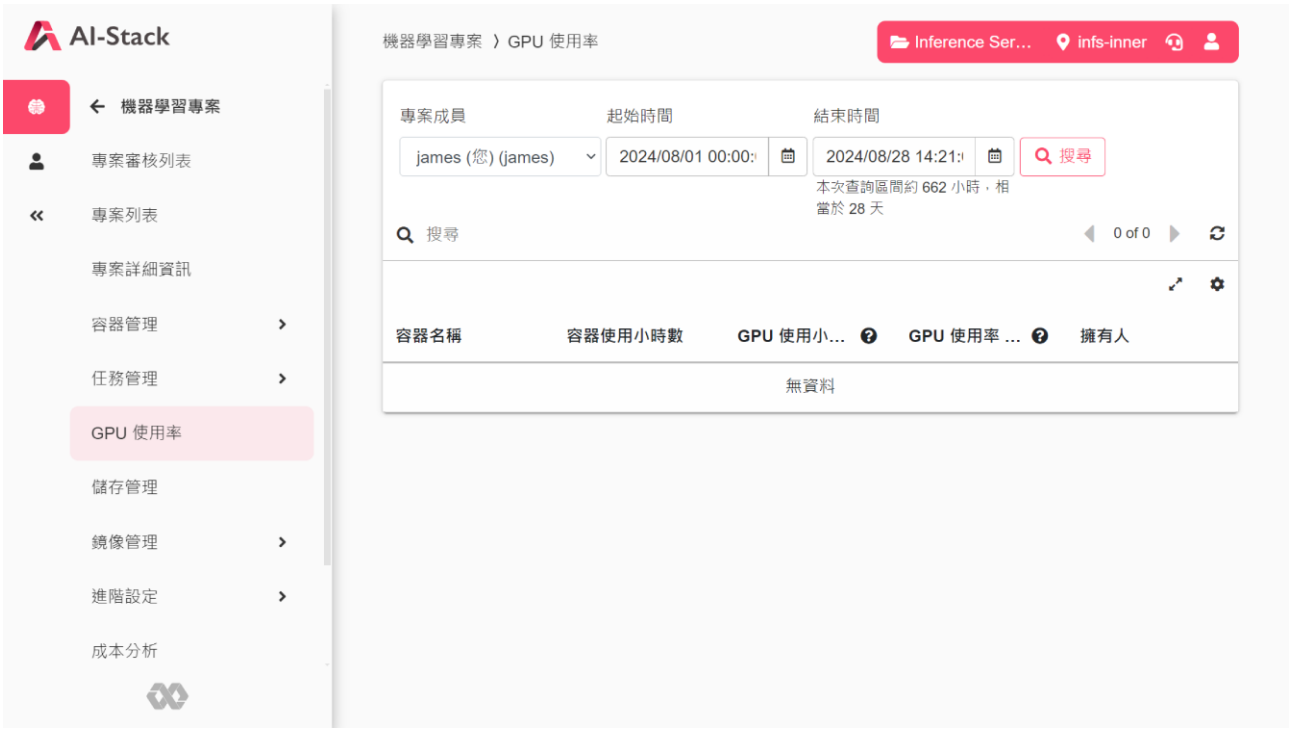


- 建立成功將出現在任務排程列表。(同樣地，任務列表中亦將新增一筆即將觸發的任務且無法直接刪除。)



5.6 GPU 使用率

【GPU 使用率】功能可查看各專案成員所開啟容器的 GPU 使用狀況，可依據容器名稱進行容器使用時數、GPU 使用時數與 GPU 使用率進行統計。開啟左側選單中的【GPU 使用率】即可查看。



5.7 儲存管理

容器（Container）的檔案系統是短暫的，當容器異常重啟或被銷毀時，檔案系統中的資料會隨之丟失，此為容器化服務的特性。要在容器生命週期之外持久保存資料，可透過掛載儲存裝置（Volume）來實現。

使用者可自行建立資料儲存裝置，供建立容器時可選擇之儲存裝置及指定掛載路徑，提供使用者進行訓練時的額外存放空間。

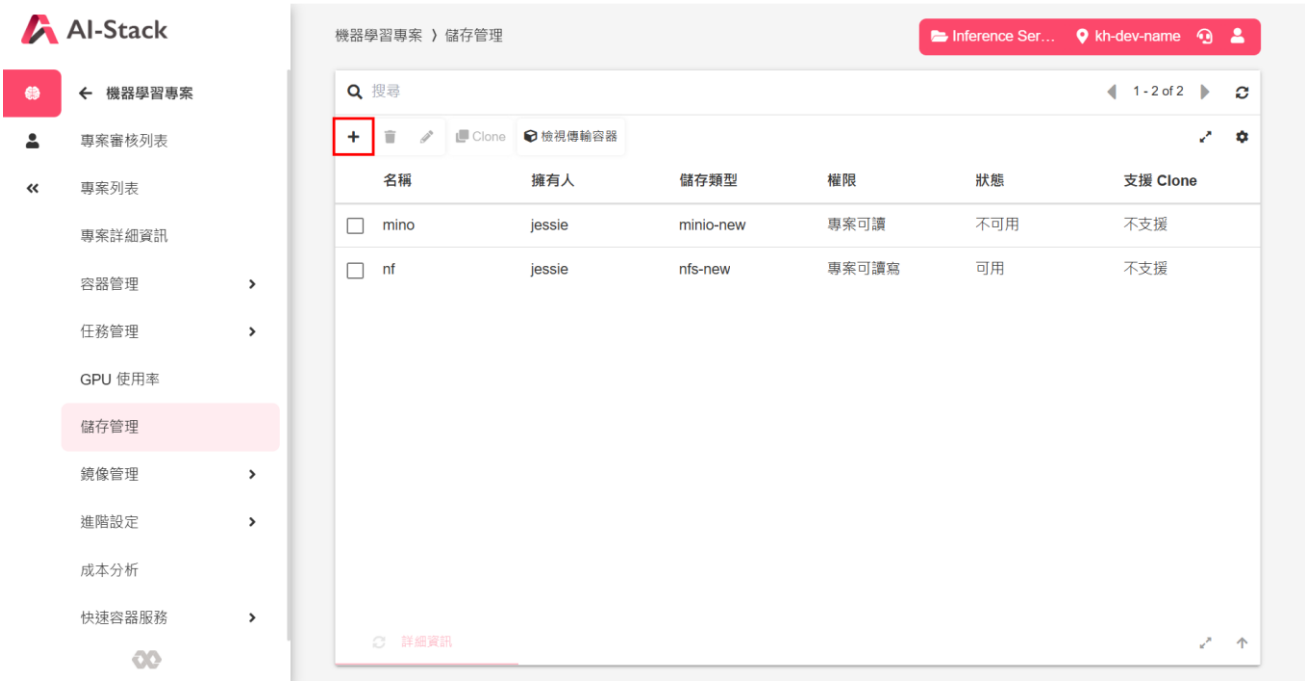
5.7.1 建立儲存裝置

平台提供的儲存裝置類型由平台管理者從後台建立儲存叢集中設定。

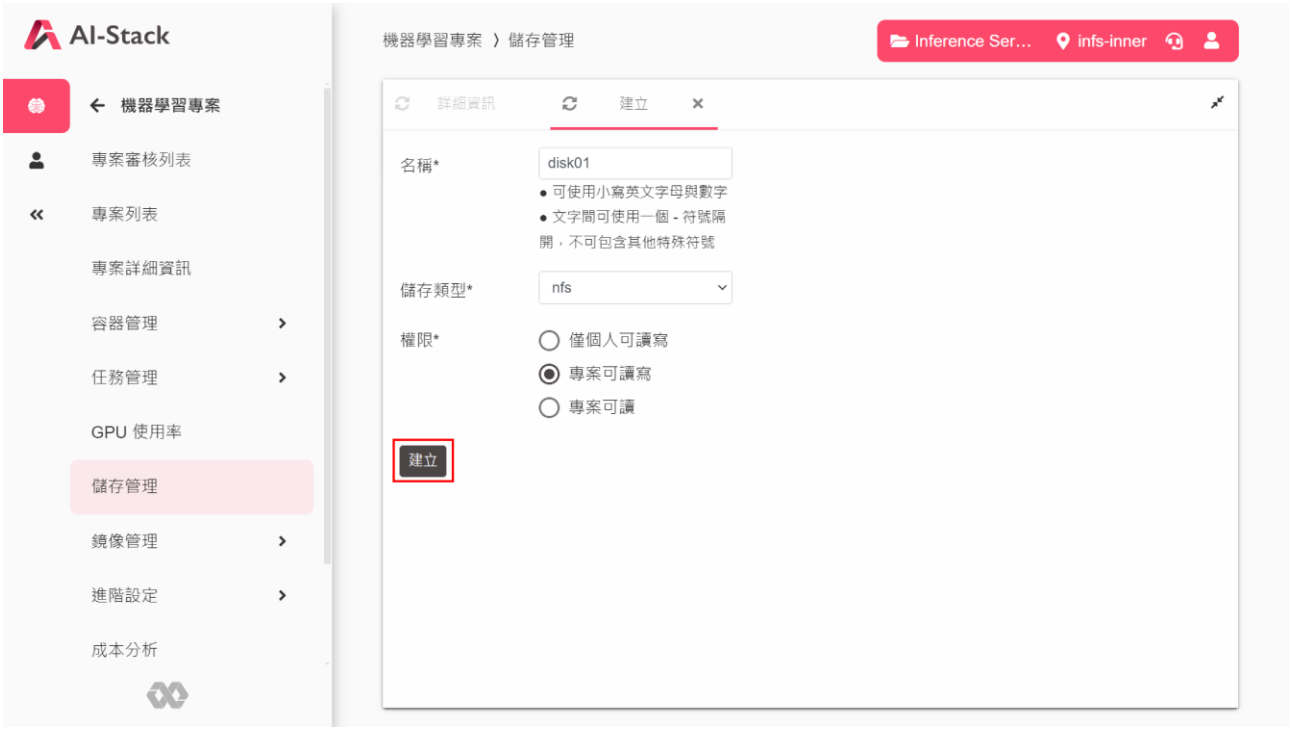
5.7.1.1 建立 NFS 儲存裝置

能將 NFS（網路檔案系統）掛載到 Pod 中，實現在刪除容器時，資料仍會被保存。

- 進入【儲存管理】頁面點擊左上角 ⊞ 建立。



- 輸入方便識別的名稱。
- 選擇儲存類型：**NFS**。
- 選擇權限後，確認資料無誤點擊 [建立]。



5.7.1.2 建立 MinIO 儲存裝置

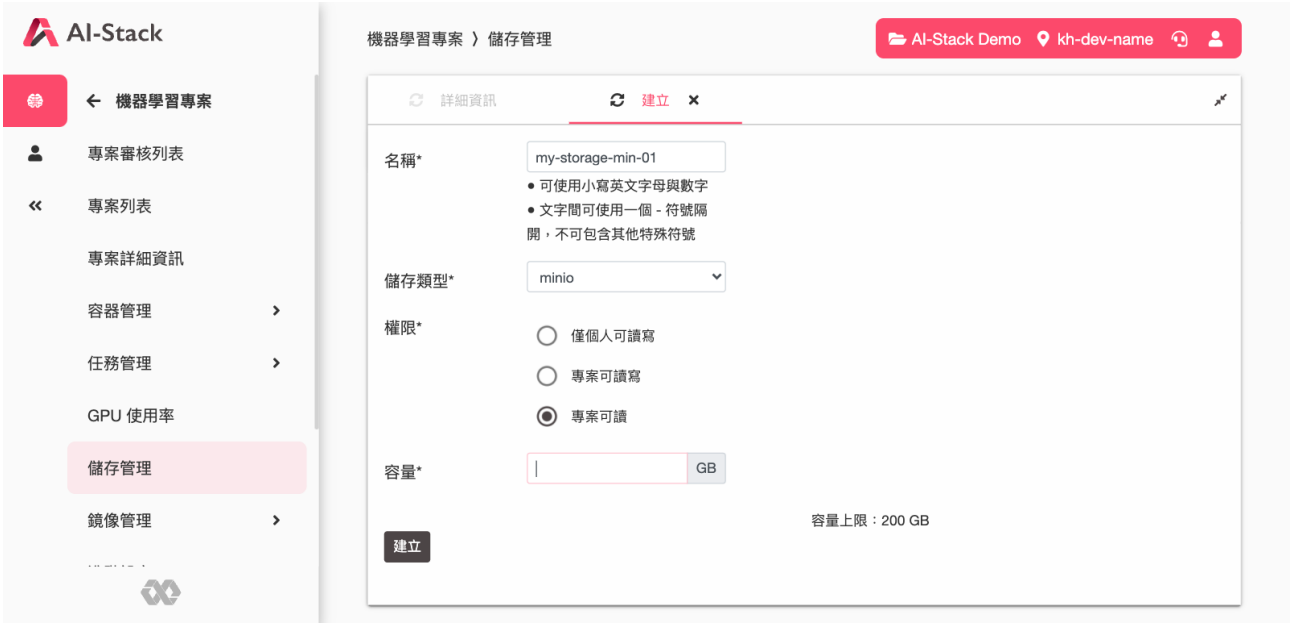
MinIO 提供高效能的物件儲存服務，適用於非結構化數據儲存，並透過 S3 兼容 API 進行存取。

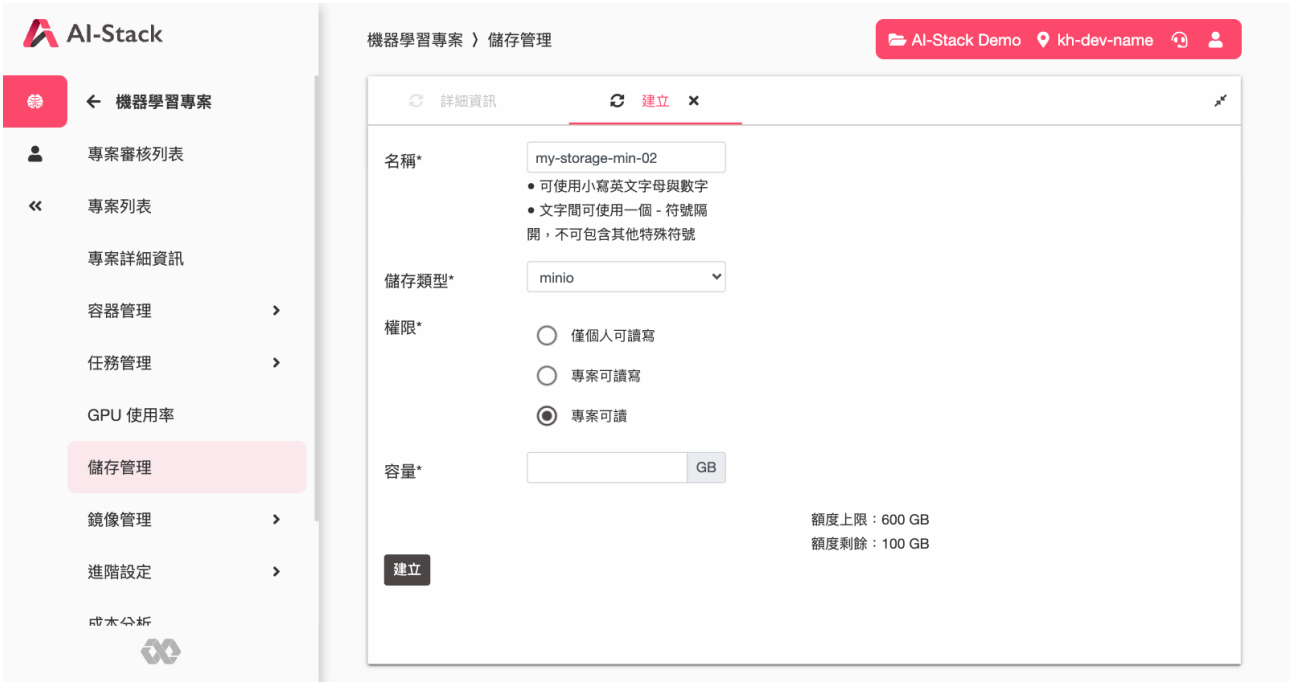
- 進入【儲存管理】頁面點擊左上角 ⊞ 建立。
- 輸入方便識別的名稱。
- 選擇儲存類型：**MinIO**。
- 設定儲存容量：選擇 MinIO 後，會出現「容量」欄位，輸入該儲存裝置的容量大小（單位：GB）。

  儲存容量顯示規則與限制說明：

  - 專案設有儲存額度上限，專案內所有 MinIO 的儲存裝置容量加總不得超過專案額度上限。
  - 專案內每個儲存裝置的容量均不得超過**單一儲存裝置容量上限**，容量上限依後台管理者設定。
  - 若**專案剩餘儲存額度**大於**單一儲存裝置的容量上限**，則系統僅顯示可設定的最大容量上限。
  - 若**專案剩餘儲存額度**小於**單一儲存裝置的容量上限**，或**未限制單一儲存容量上限**，則系統將顯示：**專案總額度上限**、**專案目前剩餘可用額度**。

- 選擇權限後，確認資料無誤點擊 [建立]，系統將同步調用 MinIO API，自動建立對應額度的 Bucket，以確保儲存設定生效。

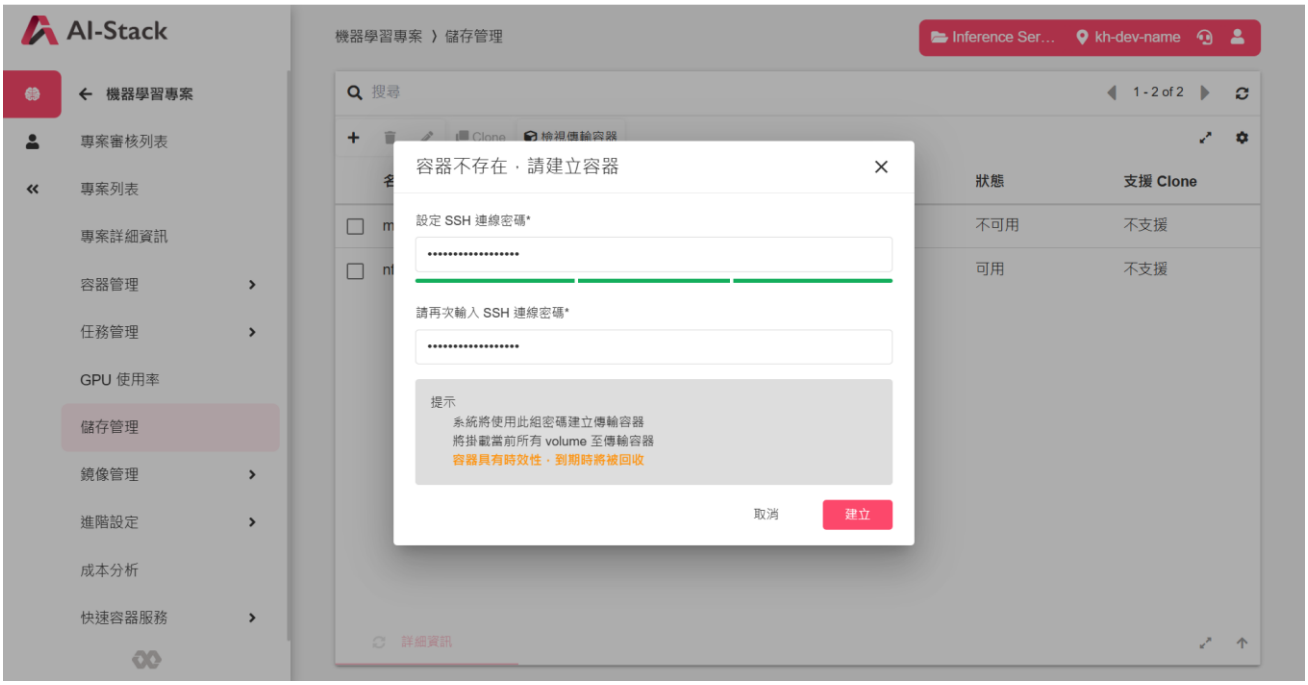




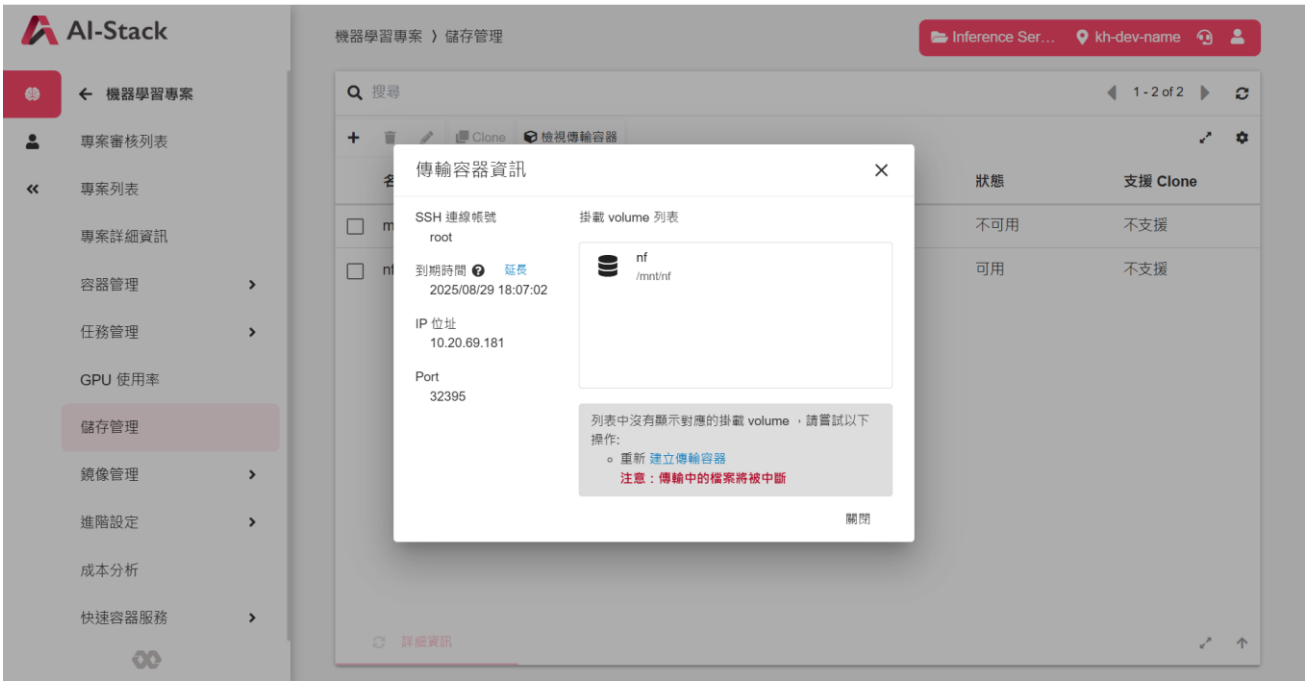
5.7.2 建立與檢視傳輸容器

可以在【儲存管理】頁面中點擊 [檢視傳輸容器]，將操作此功能的使用者當下具有存取權限的所有儲存裝置掛載，則可透過 SSH 方式進行資料的上傳、下載。

- 設定 SSH 連線密碼。



- 確認資料無誤點擊 [建立]，稍待後即可看到 [傳輸容器資訊]。

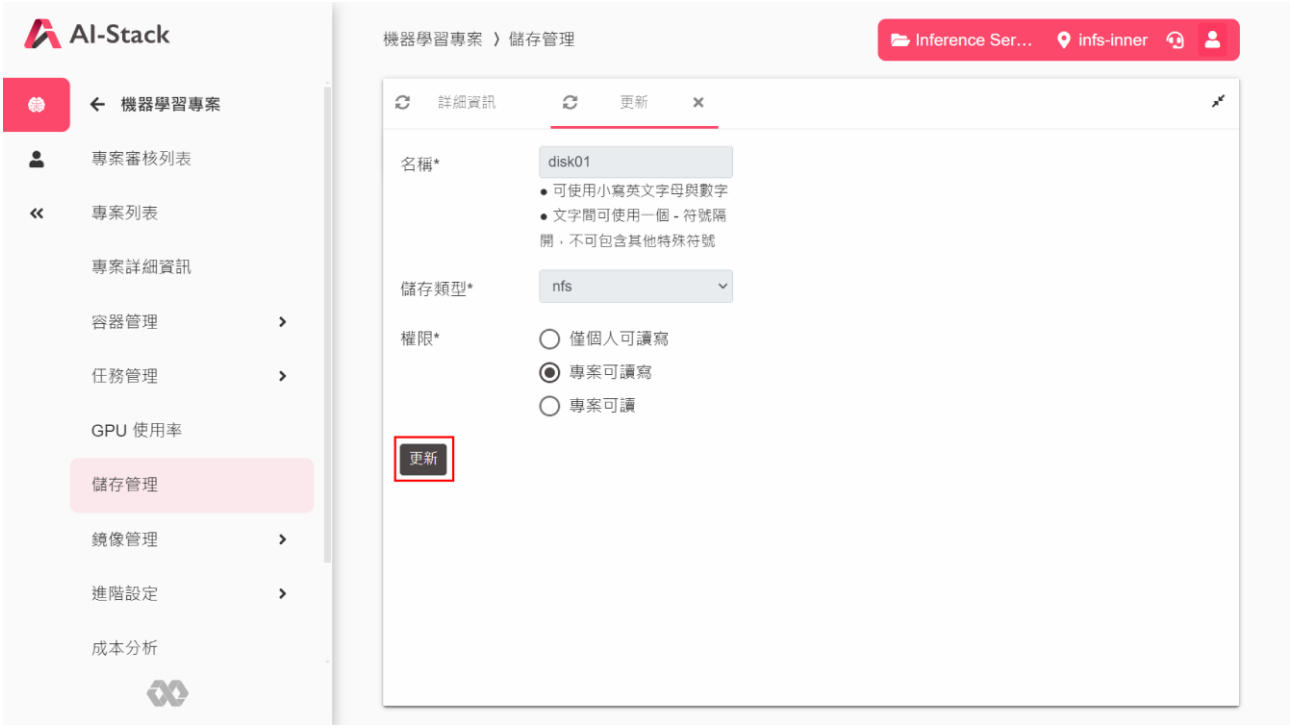


5.7.3 管理儲存裝置

5.7.3.1 更新 NFS 儲存裝置

可更新儲存裝置的讀取權限設定。

- 於清單中勾選目標儲存裝置，選定後點擊  ，可於下方看到 [更新] 頁籤。
- 修改權限並確認更新內容無誤後，點擊 [更新]。



5.7.3.2 更新 MinIO 儲存裝置

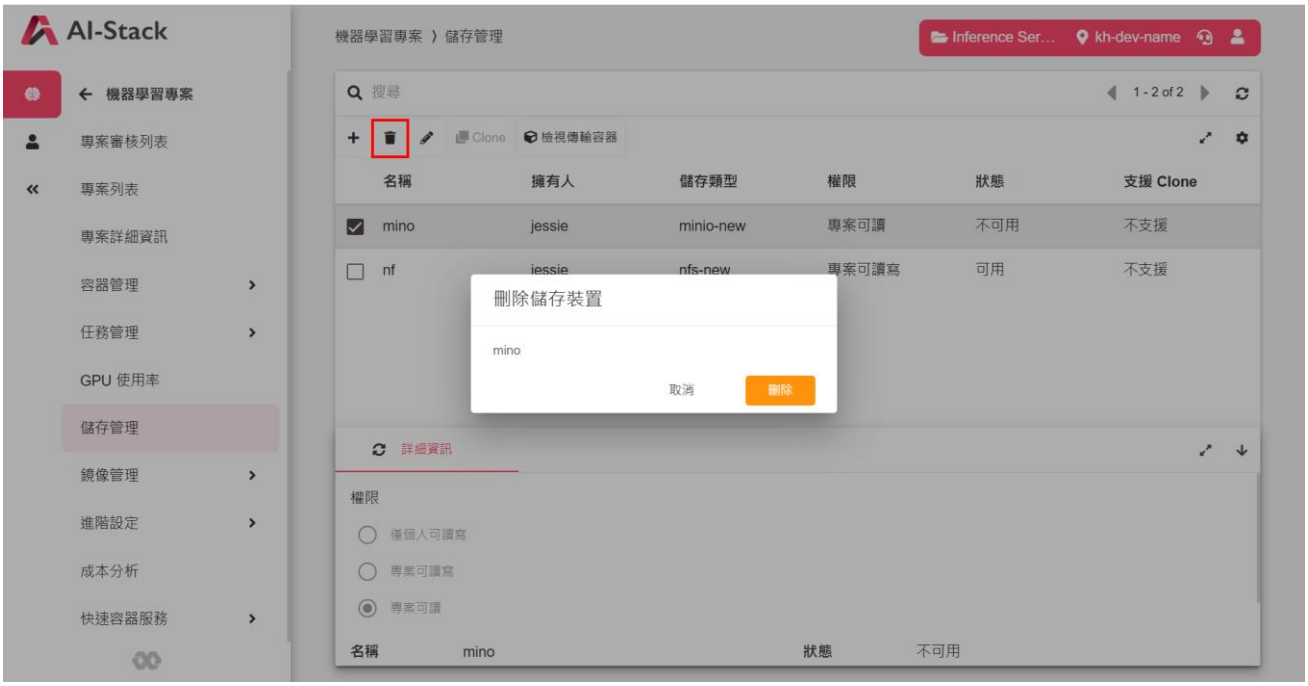
可更新儲存裝置的讀取權限以及手動擴充容量。

- 於清單中勾選目標儲存裝置，選定後點擊  ，可於下方看到 [更新] 頁籤
- 啟用 [擴充容量]，並填入 [擴充容量至] 欄位 (單位：GB)。

  * 注意：只能擴增為大於當前的容量，不支援縮減容量。
- 確認更新內容後無誤後，點擊 [更新]，更新成功後，當前儲存對應的 Bucket 會擴充至設定的額度。

5.7.4 刪除儲存裝置

- 欲刪除儲存裝置時，可於清單中勾選目標裝置，選定後點擊  將出現確認畫面，如下圖所示，確認為想要刪除的裝置後再點擊 [刪除]。



5.8 鏡像管理

為了提供機器學習開發者便捷的開發與訓練環境部署方式，平台基於常用作業環境、開發框架及 GPU 運用（如 CUDA、ROCm），預載適當的容器訪問途徑（如 JupyterLab、SSH、WebTerminal 等）。經由管理者事先在後台完成設定後，將機器學習開發服務樣板上架至公用鏡像列表，供使用者選用，省去使用者在建立容器環境時諸多繁瑣的設定。

此外，本平台提供自定義鏡像功能，使用者因個人開發需求，透過公用鏡像建立容器後，若有額外安裝套件或調整環境參數設定等情況，可透過建立自定義鏡像，將調整後的容器環境保留作為樣板，供未來或專案成員使用，免去當容器刪除重建，需要花時間重新安裝套件及調整之不便。使用者建立的自定義鏡像可於自定義鏡像列表進行檢索、編輯與刪除。

5.8.1 公用鏡像列表

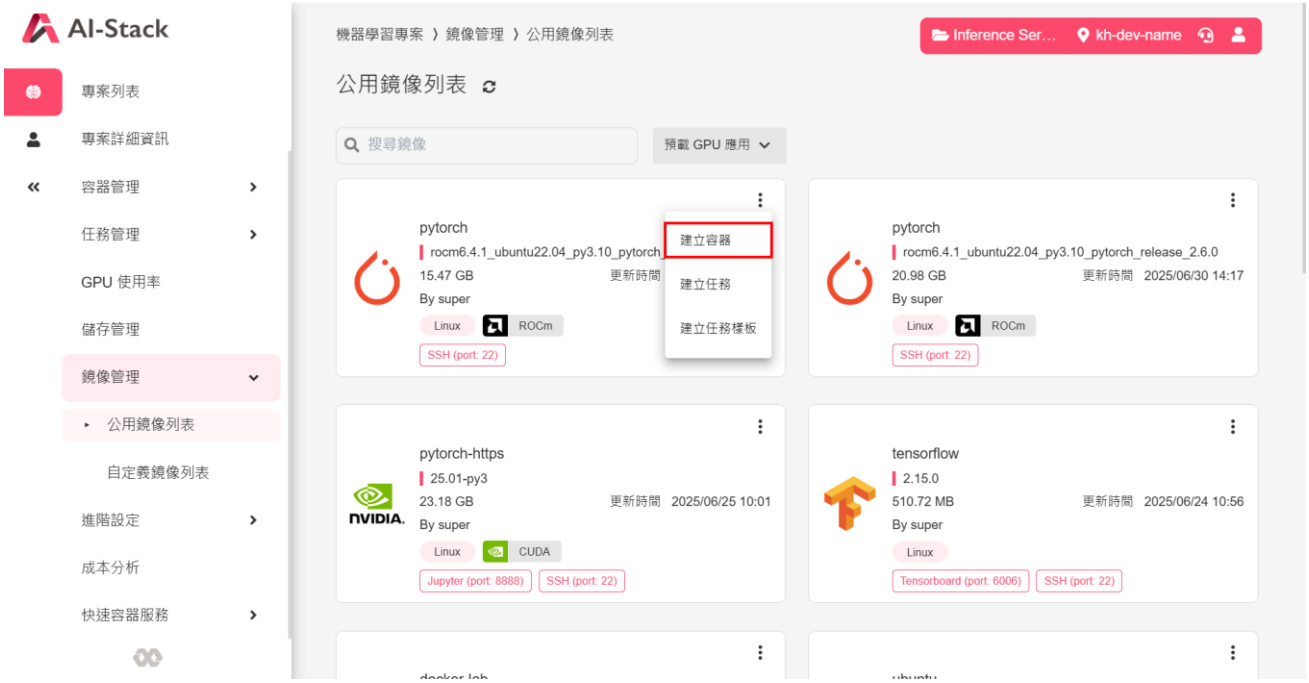
【公用鏡像列表】不僅可供查看由後台上架的公用鏡像，並可直接透過此列表將公用鏡像用於建立容器、任務以及任務樣板。

5.8.1.1 搜尋與篩選公用鏡像

- 【公用鏡像列表】頁面上方的 [搜尋鏡像] 功能可以套用不同條件與限制，方便找到指定的鏡像。
- 透過條件篩選出的鏡像結果會於下方以卡片的方式顯示。

5.8.1.2 透過公用鏡像列表建立容器

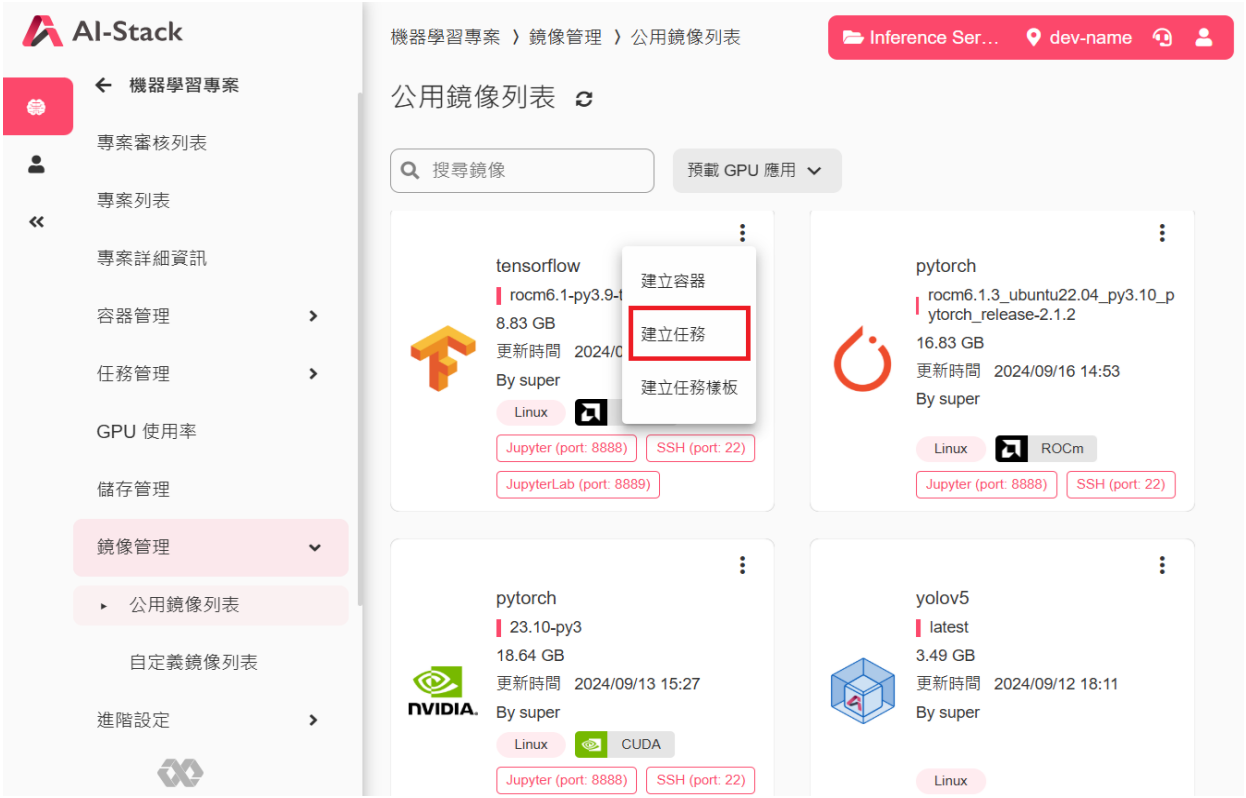
- 在【公用鏡像列表】頁面，點選要使用的鏡像卡片右上角 ⋮，並從下拉選單中選擇 [建立容器]。



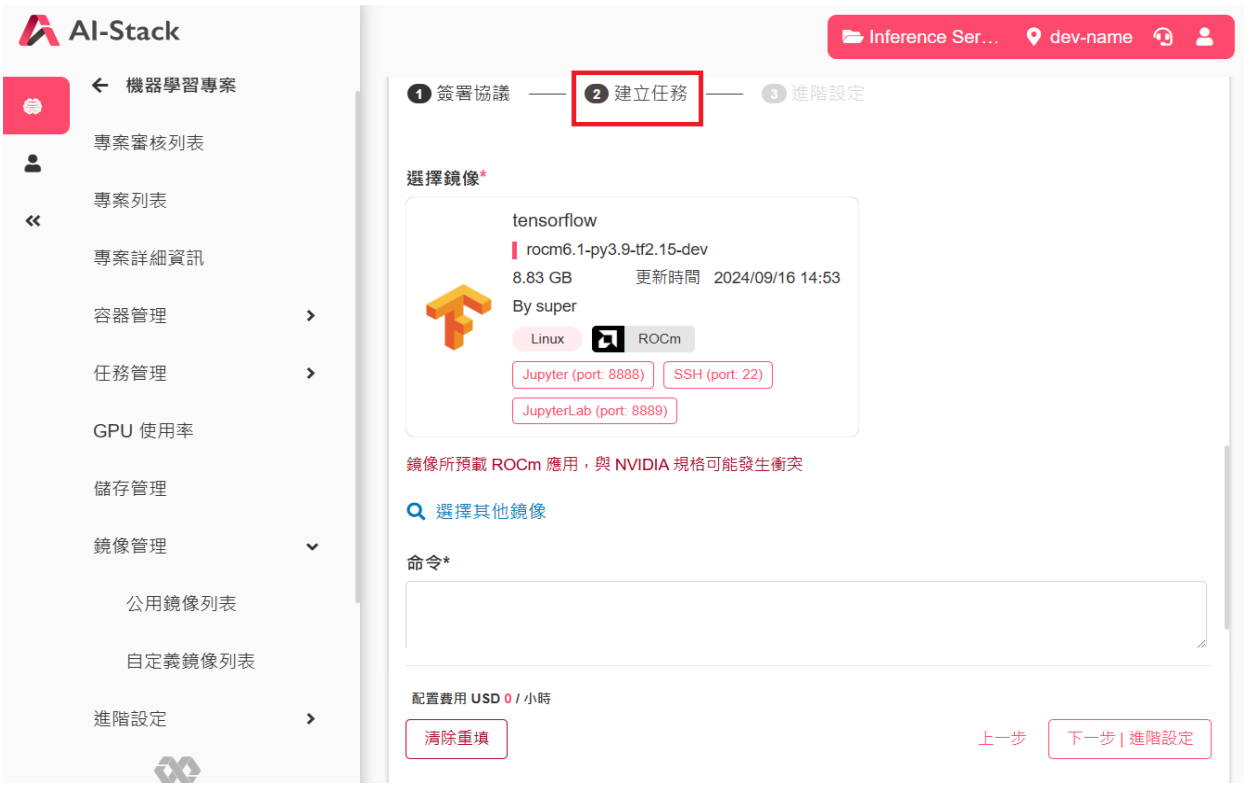
- 在 [資源配置] 的步驟中，可以看到 [鏡像] 的欄位已預選好跳轉前的鏡像。
- 也可透過下方 [選擇其他鏡像] 來更改鏡像。

5.8.1.3 透過公用鏡像列表建立任務

- 在【公用鏡像列表】頁面，點選要使用的鏡像卡片右上角 ⋮，並從下拉選單中選擇 [建立任務]。

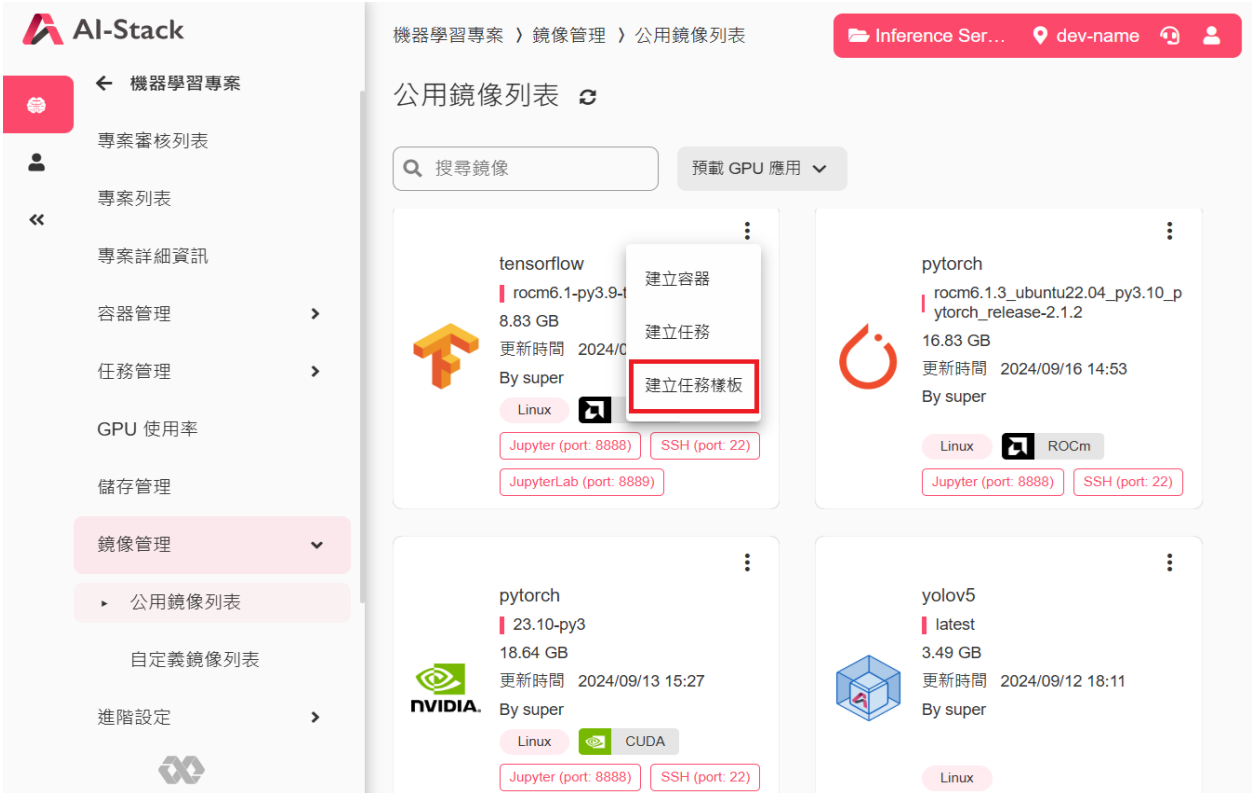


- 在 [建立任務] 的步驟中，可以看到 [選擇鏡像] 的欄位已預選好跳轉前的鏡像。
- 也可透過下方 [選擇其他鏡像] 來更改鏡像。

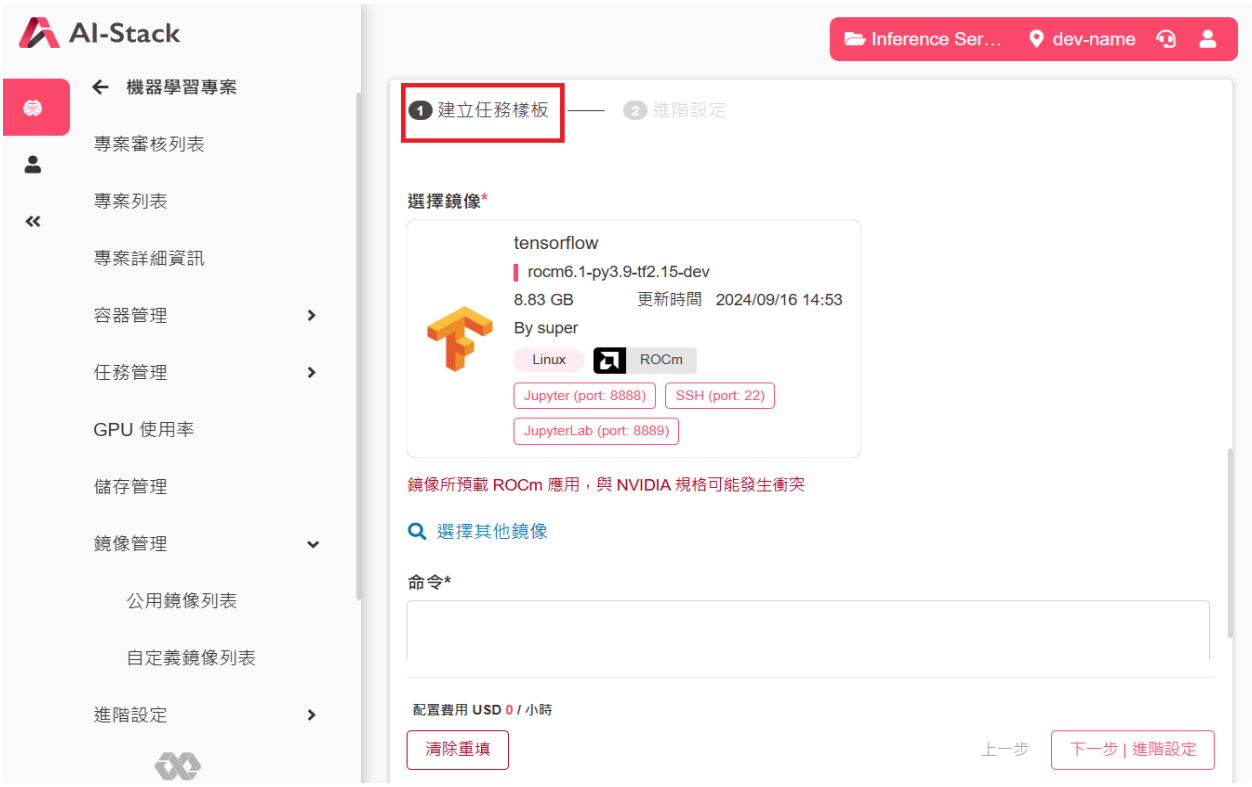


5.8.1.4 透過公用鏡像列表建立任務樣板

- 在【公用鏡像列表】頁面，點選要使用的鏡像卡片右上角 ⋮，並從下拉選單中選擇 [建立任務樣板]。



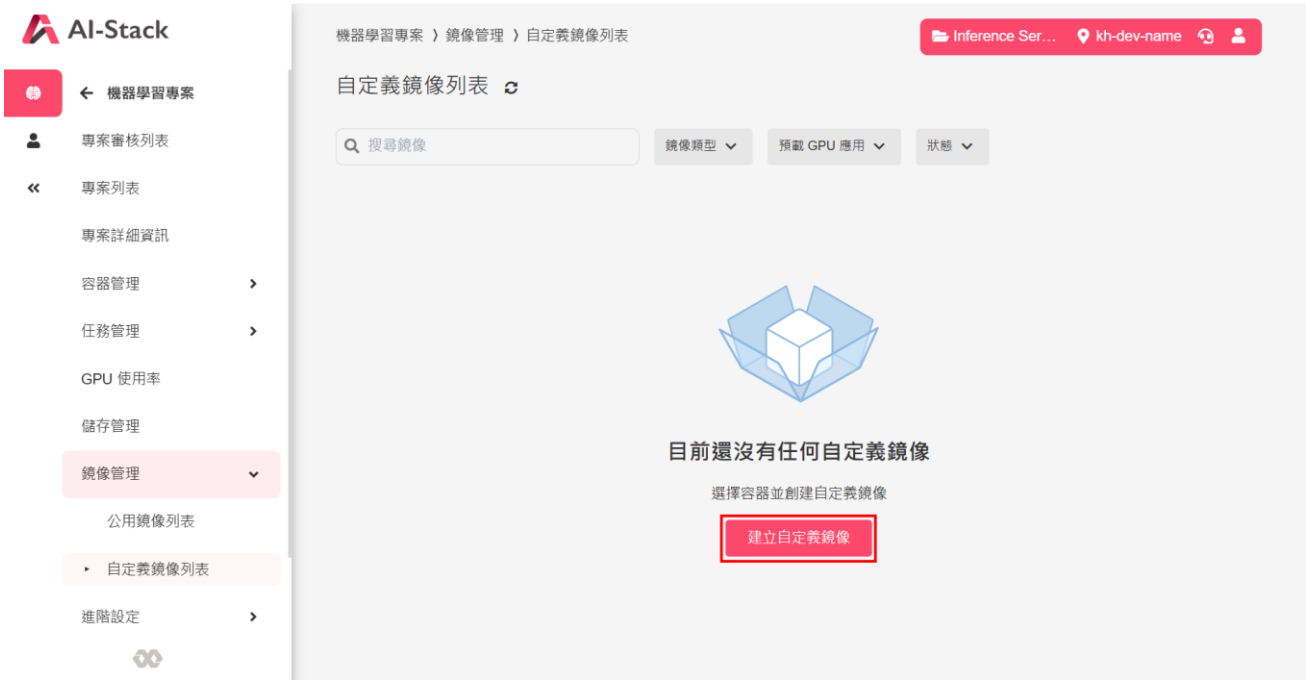
- 在 [建立任務樣板] 的步驟中，可看到 [選擇鏡像] 的欄位已預選好跳轉前的鏡像。
- 也可透過下方 [選擇其他鏡像] 來更改鏡像。



5.8.2 自定義鏡像列表

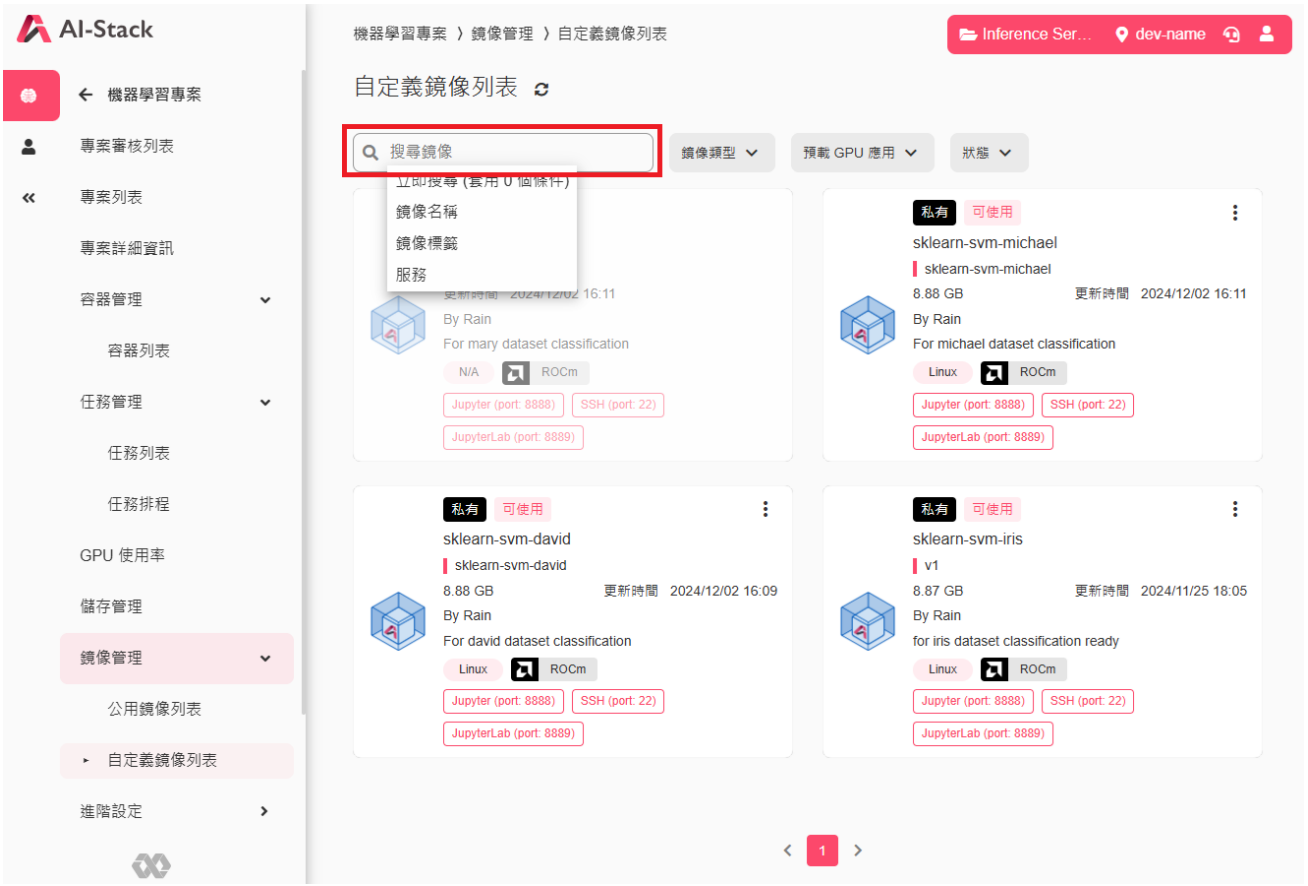
【自定義鏡像列表】列表不僅可供查看使用者自行建立的自定義鏡像，還可直接透過此列表將自定義鏡像應用於容器、任務以及任務樣板的建立。

* 注意：若先前無任何自定義鏡像，需點選畫面中 [建立自定義鏡像] 創建或選擇容器，參考建立自定義鏡像。



5.8.2.1 搜尋與篩選自定義鏡像

- 【自定義鏡像列表】頁面上方的 [搜尋鏡像] 功能可以套用不同條件與限制，方便找到指定的鏡像。
- 透過條件篩選出的鏡像結果會於下方以卡片的方式顯示。



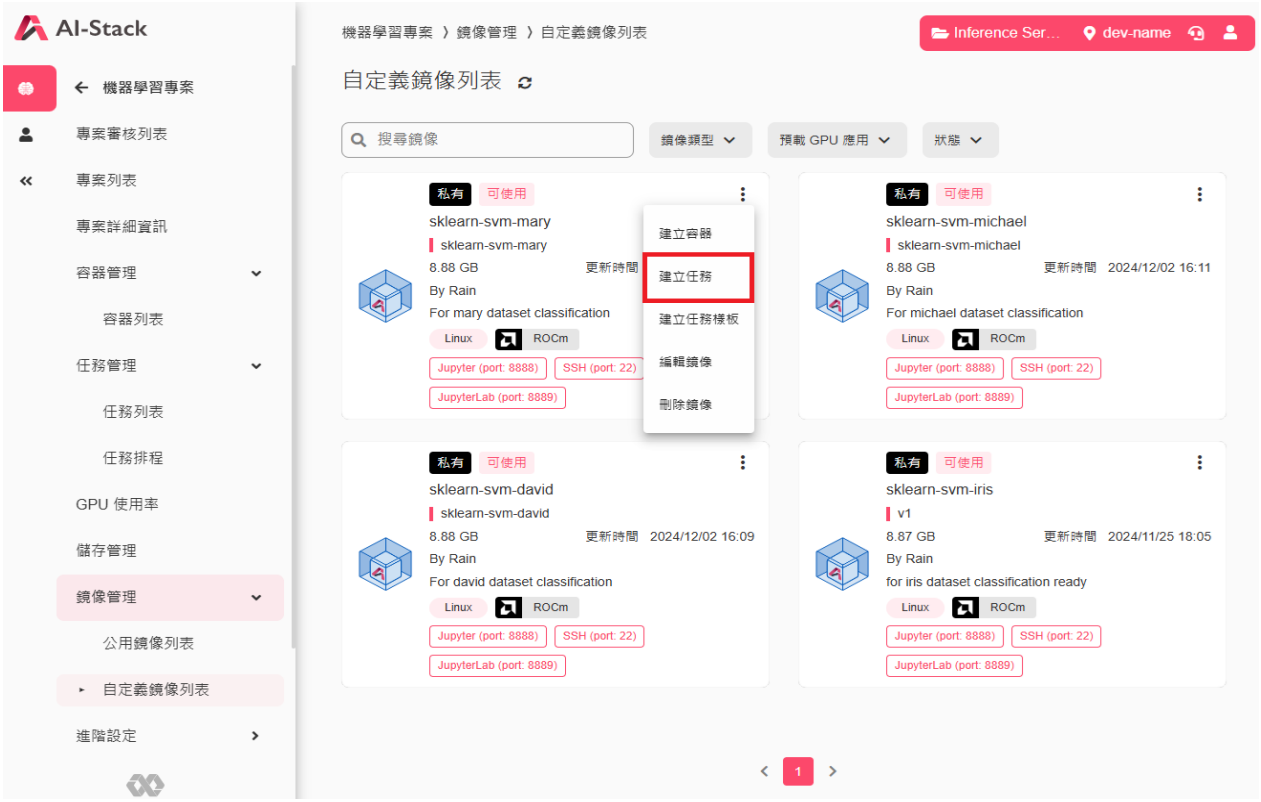
5.8.2.2 透過自定義鏡像列表建立容器

- 在【自定義鏡像列表】頁面，點選要使用的鏡像卡片右上角 ⋮，並從下拉選單中選擇 [建立容器]。

- 在 [資源配置] 的步驟中，可以看到 [鏡像] 的欄位已預選好跳轉前的鏡像。
- 也可透過下方 [選擇其他鏡像] 來更改鏡像。

5.8.2.3 透過自定義鏡像列表建立任務

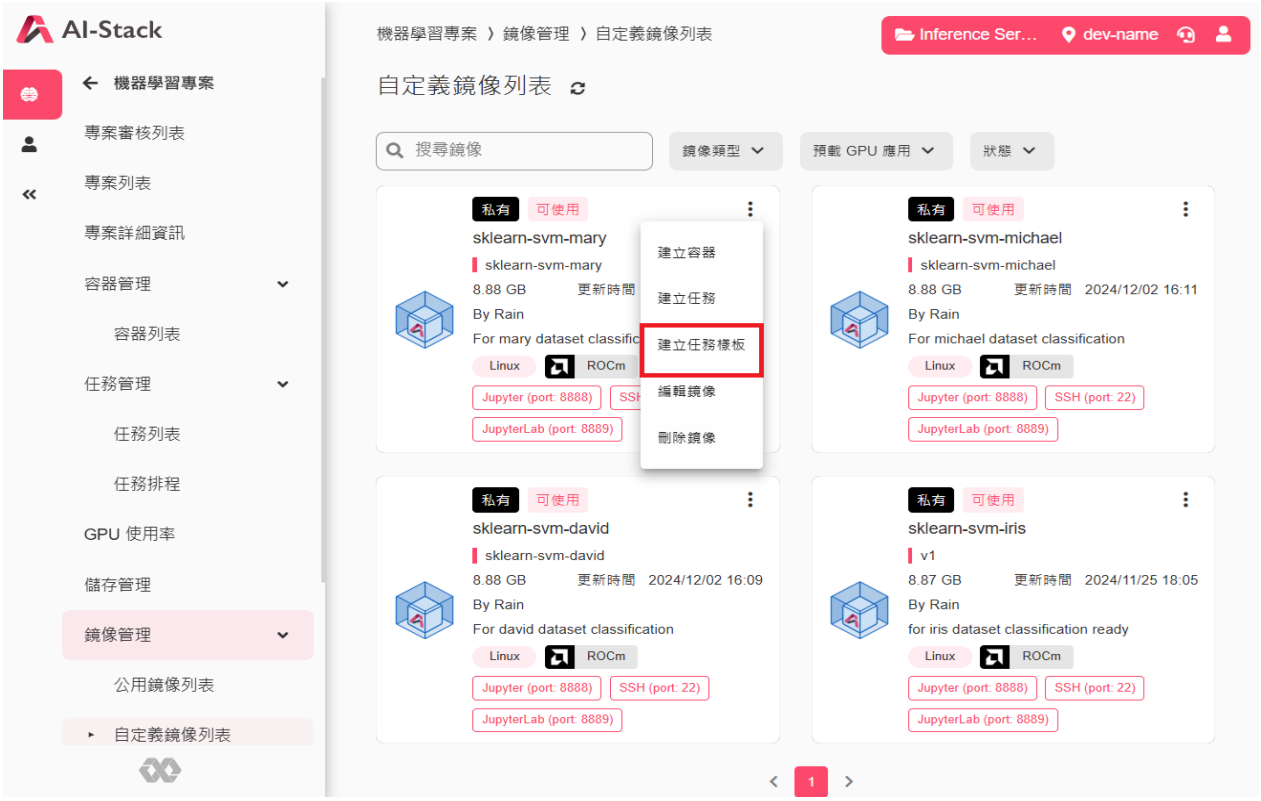
- 在【自定義鏡像列表】頁面，點選要使用的鏡像卡片右上角 ⋮，並從下拉選單中選擇 [建立任務]。



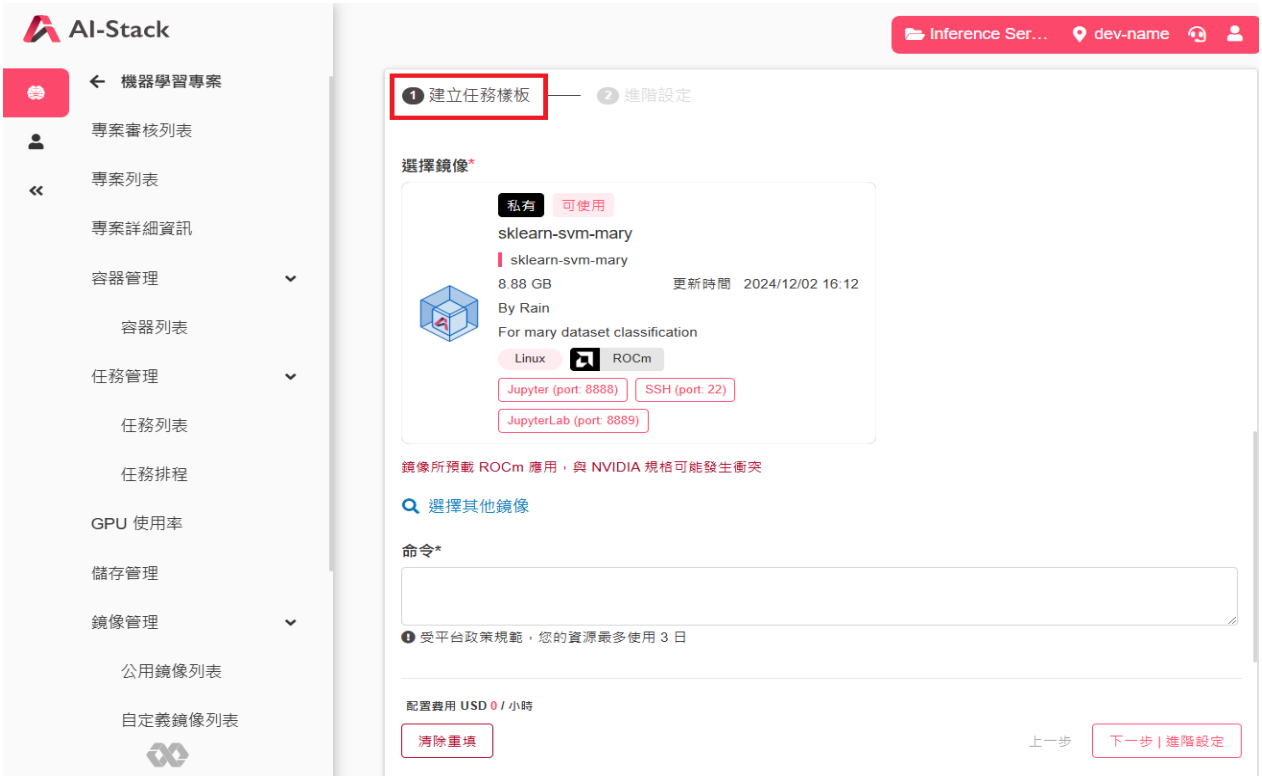
- 在 [建立任務] 的步驟中，可以看到 [選擇鏡像] 的欄位已預選好跳轉前的鏡像。
- 也可透過下方 [選擇其他鏡像] 來更改鏡像。

5.8.2.4 透過自定義鏡像列表建立任務樣板

- 在【自定義鏡像列表】頁面，點選要使用的鏡像卡片右上角 ⋮，並從下拉選單中選擇 [建立任務樣板]。



- 在 [建立任務樣板] 的步驟中，可看到 [選擇鏡像] 的欄位已預選好跳轉前的鏡像。
- 也可透過下方 [選擇其他鏡像] 來更改鏡像。

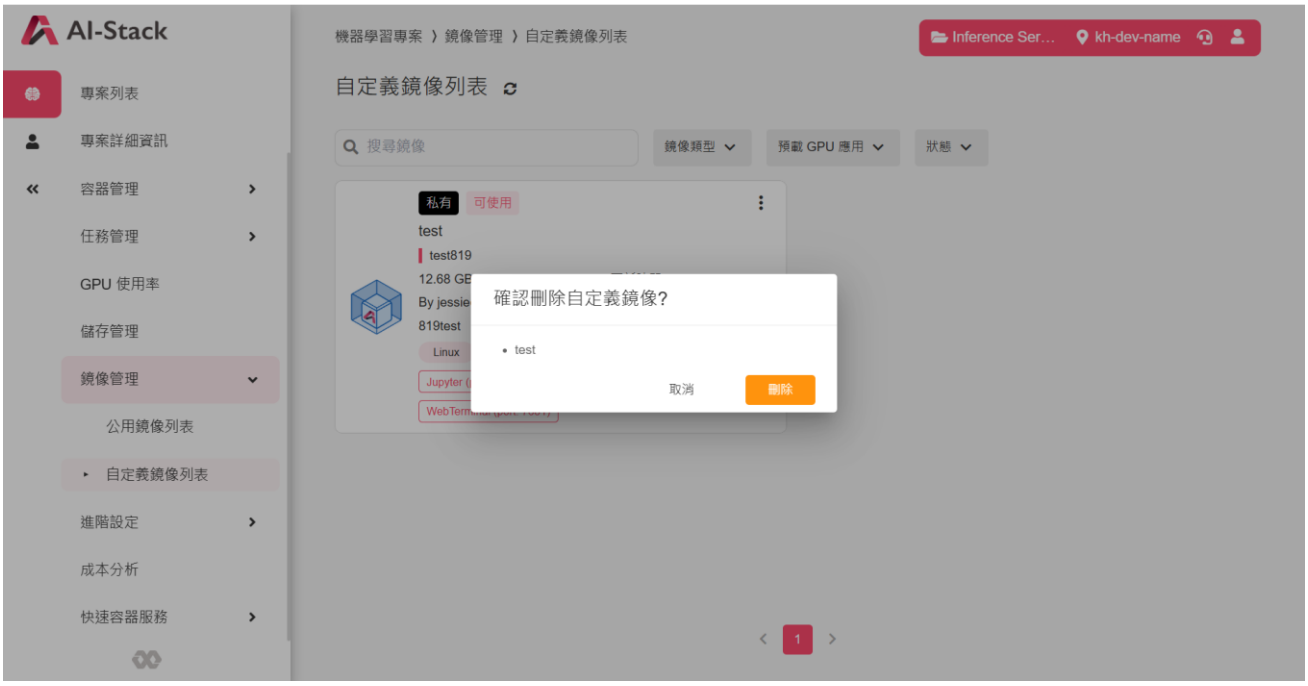


5.8.2.5 編輯自定義鏡像

- 在【自定義鏡像列表】頁面，點選要使用的鏡像卡片右上角 ⋮，並從下拉選單中選擇 [編輯鏡像]，會跳轉到有關鏡像的設置，確認更改後點擊 [儲存]。

5.8.2.6 刪除自定義鏡像

- 在【自定義鏡像列表】頁面，點選要使用的鏡像卡片右上角 ⋮，並從下拉選單中選擇 [刪除鏡像]，隨後會彈出視窗詢問是否確認刪除該鏡像，確認後點擊 [刪除] 完成操作。



5.9 進階設定

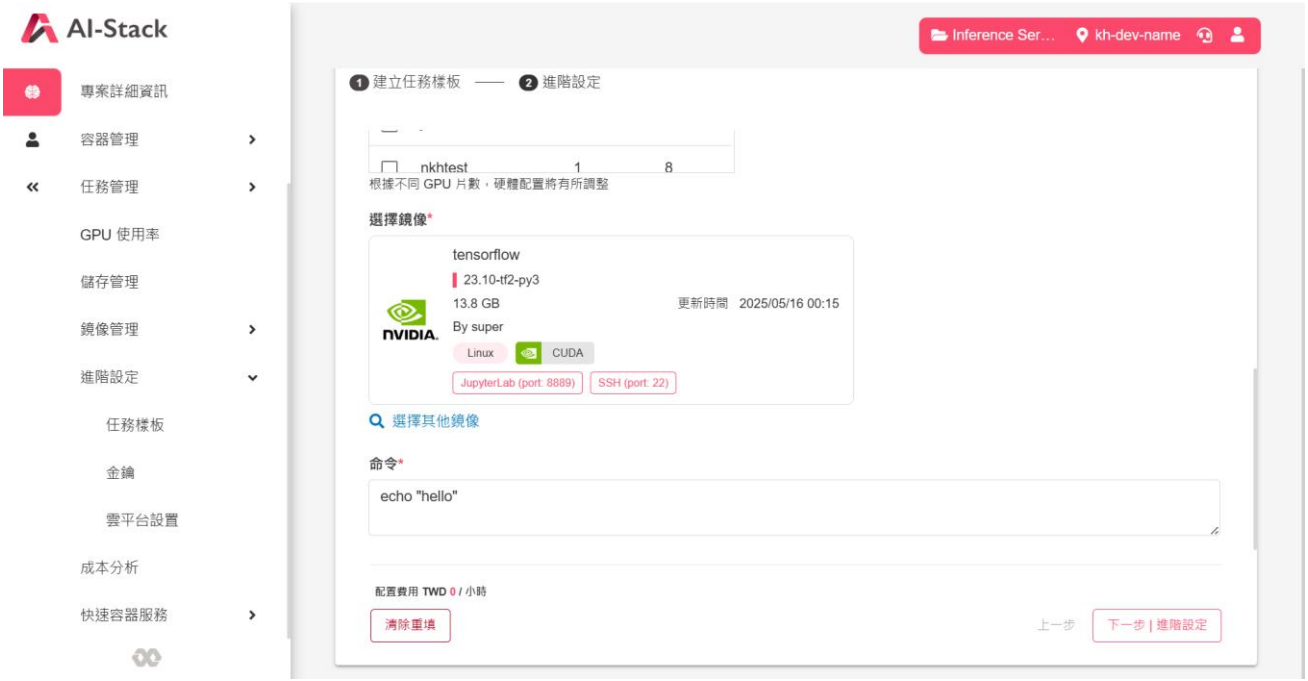
5.9.1 任務樣板

設置任務排程的其中一種方式，為透過預先建立的任務樣板進行排程設定（可參考任務排程一節所述），且透過預先建立好的任務樣板，亦可以直接啟動執行。

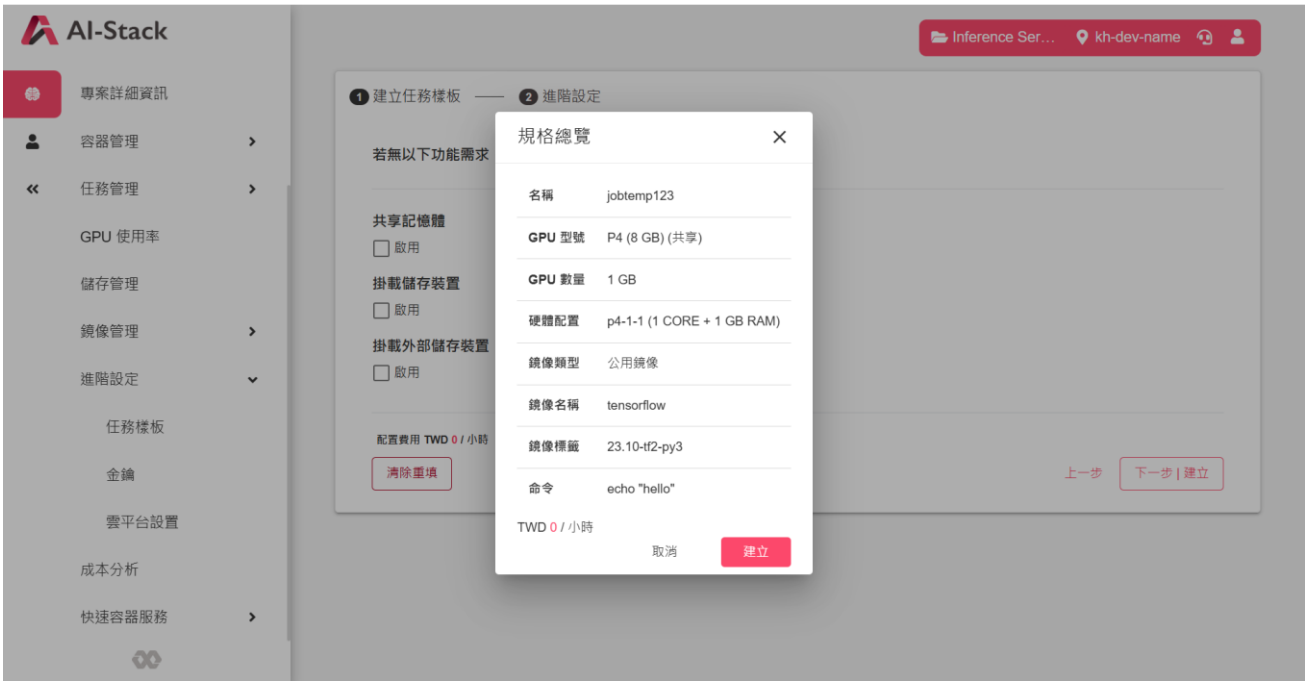
5.9.1.1 建立任務樣板

- 從左側選單開啟【進階設定】>【任務樣板】。
- 點擊左上方 ⊞ 圖示進入建立任務樣板頁面。
- 填入名稱、選擇 GPU 型號、GPU 數量、硬體配置。

- 選擇鏡像並輸入命令。
- 點擊 [下一步 | 進階設定]。



- 若不需啟用進階設定的功能，可直接點擊 [下一步 | 建立] 跳出規格總覽頁面。
- 於規格總覽頁面確認內容無誤後可點擊 [建立]，新增成功會導到任務樣板頁面。

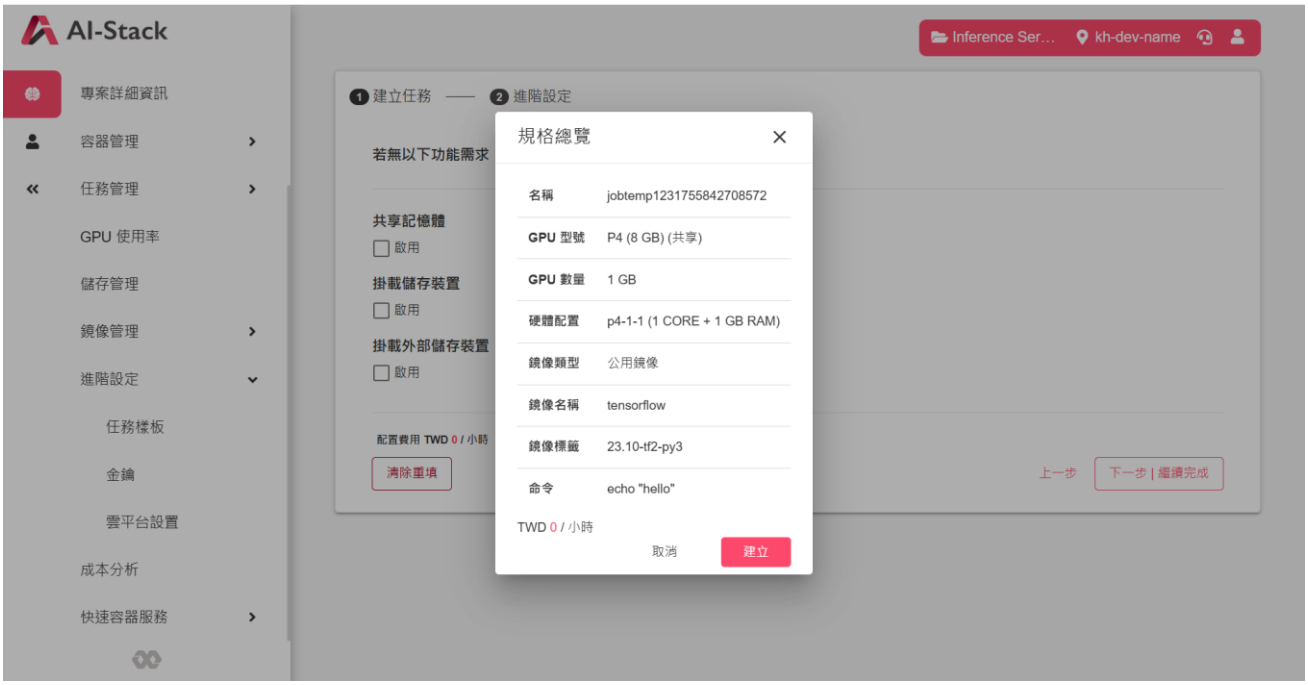


5.9.1.2 執行任務樣板

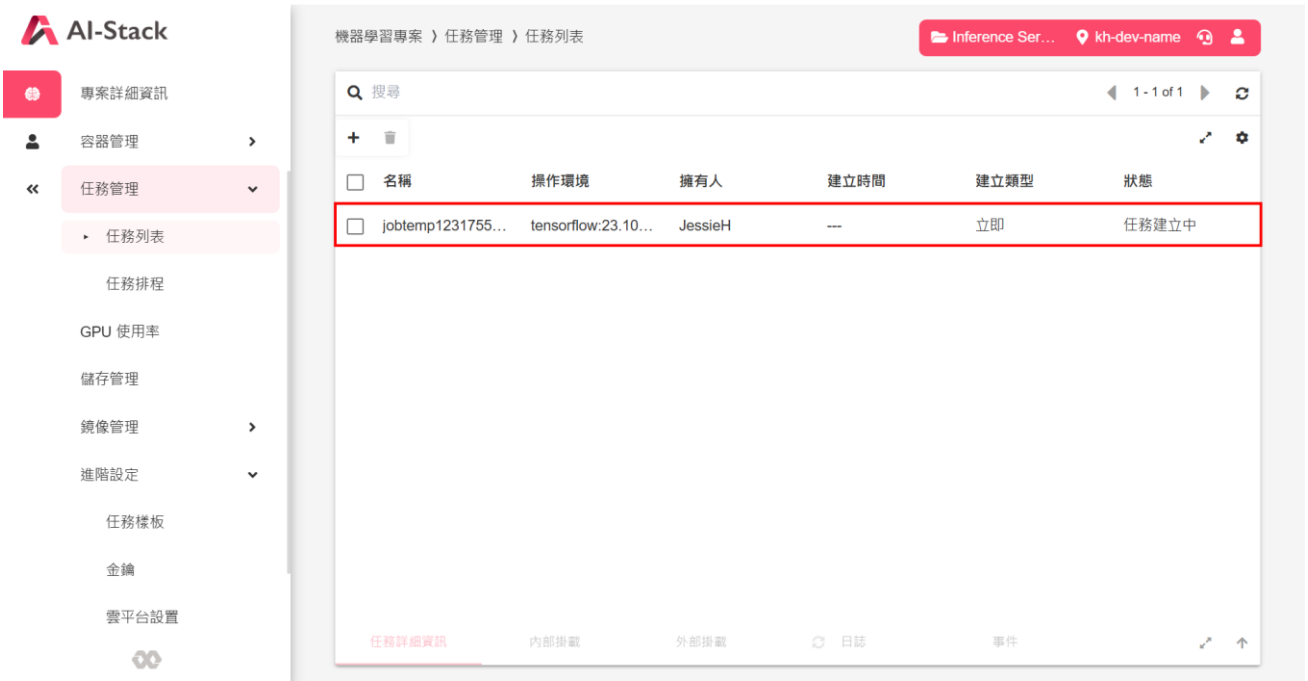
- 勾選一筆欲執行的任務樣板，點擊上方 [執行任務]。

- 將帶入樣板內容並開啟任務頁面，名稱一欄預設使用樣板名稱加序號之組合，如有需要此內容可自行修改。
- 點擊 [下一步 | 進階設定]。

- 若不需啟用進階設定的功能，可直接點擊 [下一步 | 繼續完成] 跳出規格總覽頁面。
- 於規格總覽頁面確認內容無誤後可點擊 [建立]。



- 建立完成將跳轉【任務列表】，會看到多出一筆狀態為 [任務建立中] 的任務，且將立即執行。



5.9.2 金鑰

提供使用者統一管理管理金鑰之功能。

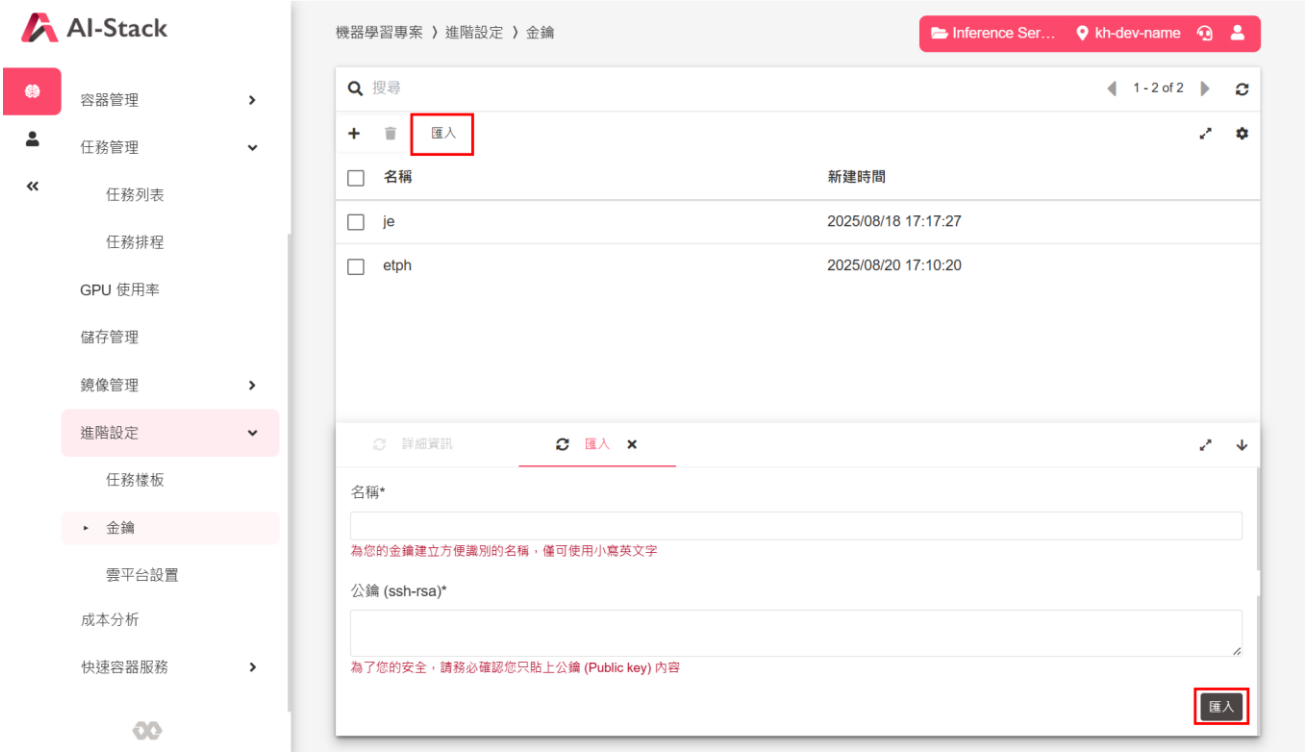
5.9.2.1 新增金鑰

使用者新增之金鑰可供建立機器學習服務時做為 SSH 連線使用，開啟左側選單【進階設定】>【金鑰】頁面，按下左上方 ⊞ 圖示建立新的金鑰，並下載私鑰 * 至使用者電腦中。

* 私鑰需妥善保管，若遺失將亦無法找回，僅能重新建立新的金鑰，同時將導致無法連入設定使用本金鑰的服務。

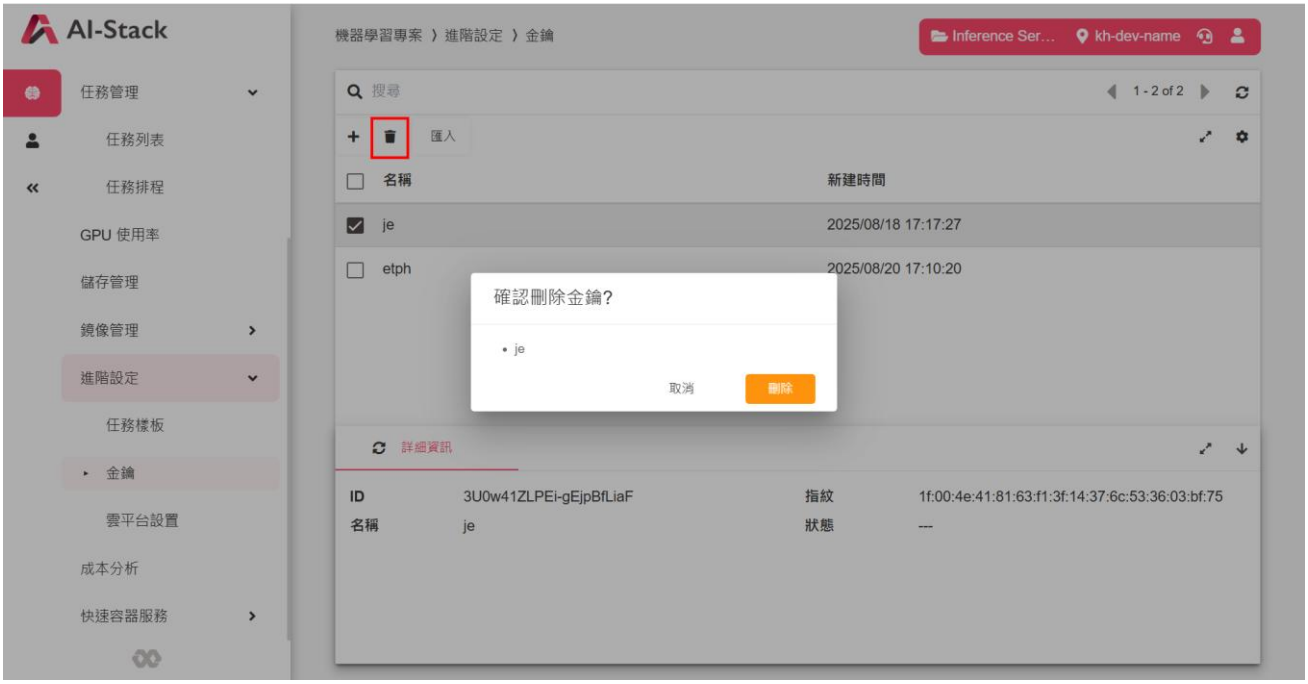
5.9.2.2 匯入金鑰

若有習慣使用的金鑰，可按下 [匯入] 將金鑰匯入系統中，填寫名稱並填入該金鑰的公鑰，請注意安全性，確保匯入的為公鑰，按下匯入即可在本系統中使用該金鑰。



5.9.2.3 刪除金鑰

欲刪除金鑰時，於清單中勾選欲刪除之金鑰，再點擊  圖示，將會出現確認畫面，如下圖所示，若確認無誤按 [刪除] 即可予以刪除。



5.10 成本分析

本平台提供以專案為主的成本統計資訊，其統計基準為各容器使用期間（以小時為單位）乘每一容器套用的 MLS 規格每小時費用 * 。

* 備註：MLS 規格及每小時費用之設定為平台管理者功能，需從管理後台操作。

5.11 快速容器服務（RCS）

AI-Stack 整合容器化技術和 Kubernetes 系統在平台中。透過建立容器的方式提供統一的運行環境，協助 AI 模型開發者將開發環境與應用程式和模組間的相互依賴全部打包，讓 AI 應用能夠在各種環境中選擇需要的版本，確保運行一致。有效進行版本控制，提高效率。再透過 Kubernetes 編排與管理容器、做自動化容器擴展，輕鬆管理大規模容器化應用，以及自動化運行，並按需求動態調整。在兩者技術及系統的協同運作下，AI-Stack 讓 AI 開發到推論應用歷程更加高效可靠，進而加速 AI 模型從實驗室到生產環境的轉化，且大大提升 AI 應用的可維護性、可擴展性，以及環境可移植性。

本節以部署一個 nginx 應用為例，說明運用「快速容器服務（RCS, Rapid Container Service）」功能產生一個透過連結點擊即用的應用。

5.11.1 部署應用程式

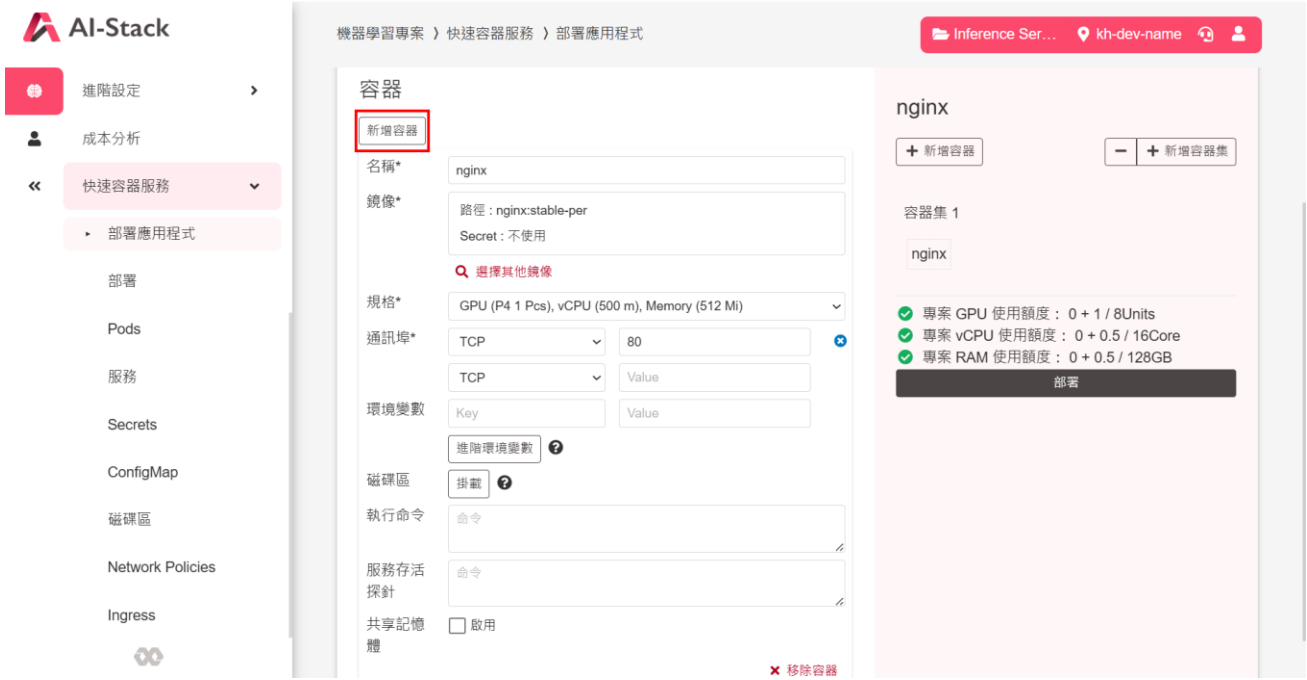
- 設定部署名稱、歷史版本紀錄上限、容器集數量。

- 點擊 [新增容器]
  - 名稱：輸入方便識別的名稱 (例如：nginx)。
  - 鏡像：可以選擇 [公用鏡像]、[自定義鏡像] 或 [手動下載]。

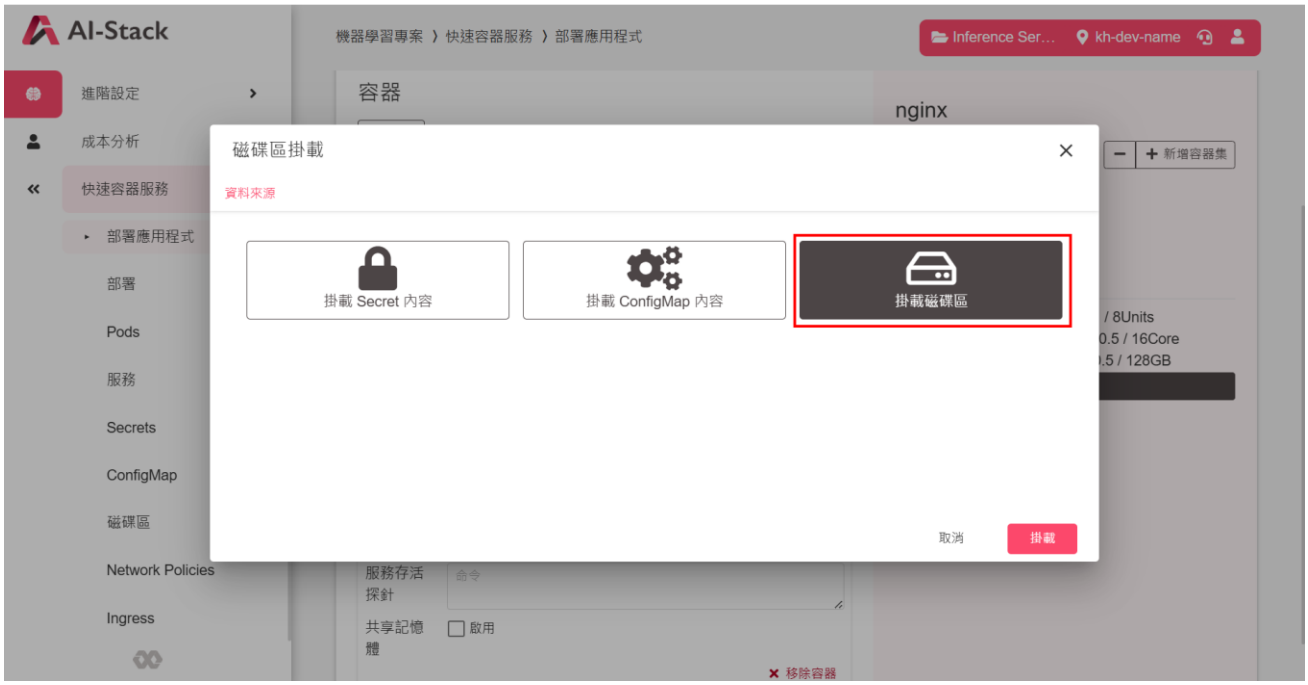
    本示範案例選擇 [手動下載]，並輸入 [鏡像路徑] (例：nginx:stable-perl)，選擇不使用 [Secret]，因鏡像存放在公共的 Docker Hub，不須要額外的登入資訊即可拉取。

    *** 備註：有關 [執行命令] 欄位**

    當從運行中的容器生成鏡像 (使用 save 或 export) 時，鏡像內的啟動指令 (例：CMD 或 ENTRYPOINT) 會丟失，導致在 RCS 中部署該鏡像時容器無法正常啟動，進而影響系統運行。為解決此問題，可在 [執行命令] 欄位手動補充或覆蓋容器啟動指令，以確保容器能正確啟動並執行預期任務。
  - 規格：選擇所需硬體規格 (本示範案例先以最低規格選取)。
  - 通訊埠：TCP 80。
  - 環境變數：不設定，本範例不使用環境變數。
  - 共享記憶體：可以選擇是否啟用共享記憶體。

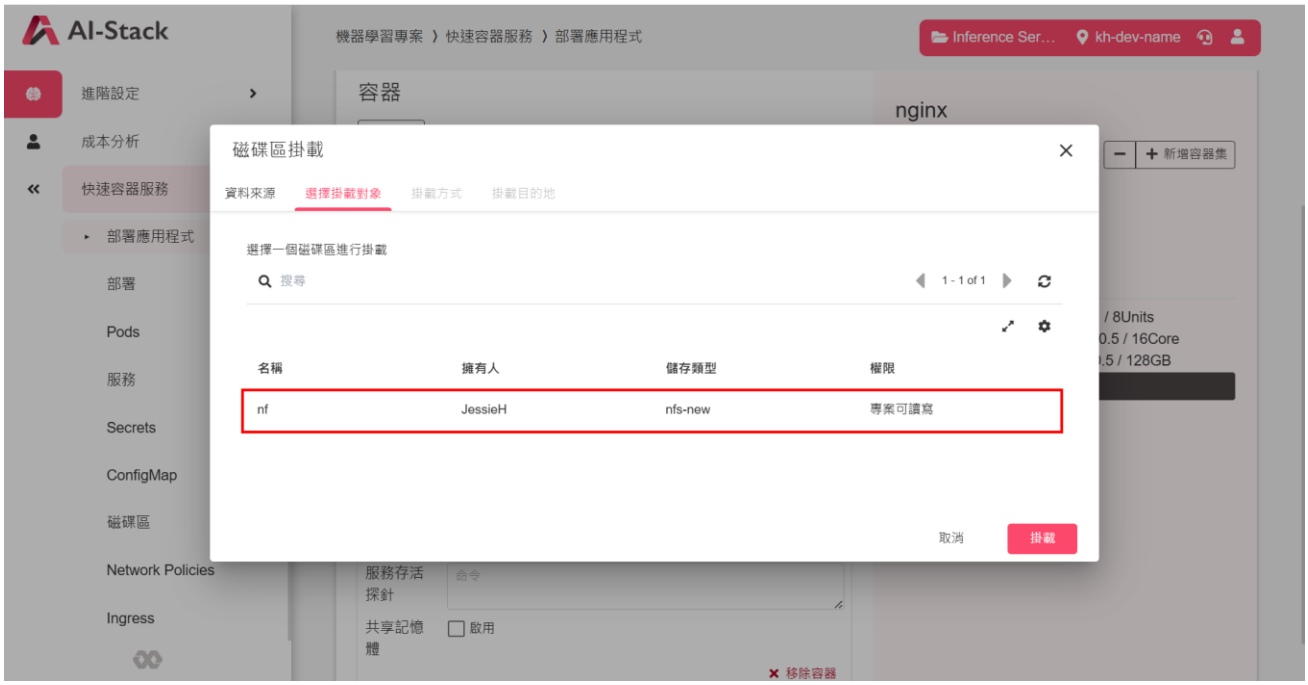


- 磁碟區：點擊 [掛載]，會出現設定視窗。
  - 點擊 [掛載磁碟區]。



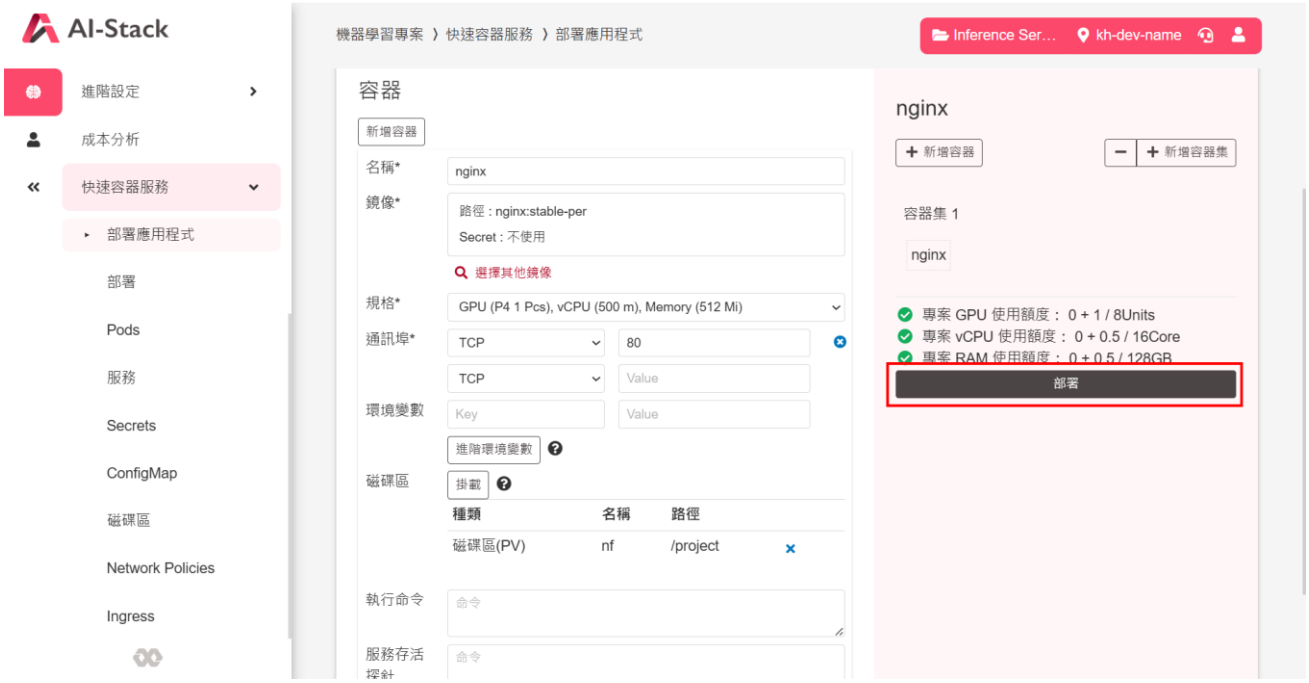
- 選擇要掛載的磁碟區 * 進行掛載。

  * 注意：若無可掛載的磁碟區，建立磁碟區進行建立。



- 在掛載路徑填入要掛載的資料夾路徑（/project）。
- 點擊 [掛載]。

- 正式部署前，可在右側可視化的介面查看。例如，由下圖可知將部署一個容器集，容器集內有一個名為 nginx 的容器，確認資料無誤，點擊 [部署] 後，接續管理部署。



5.11.2 部署

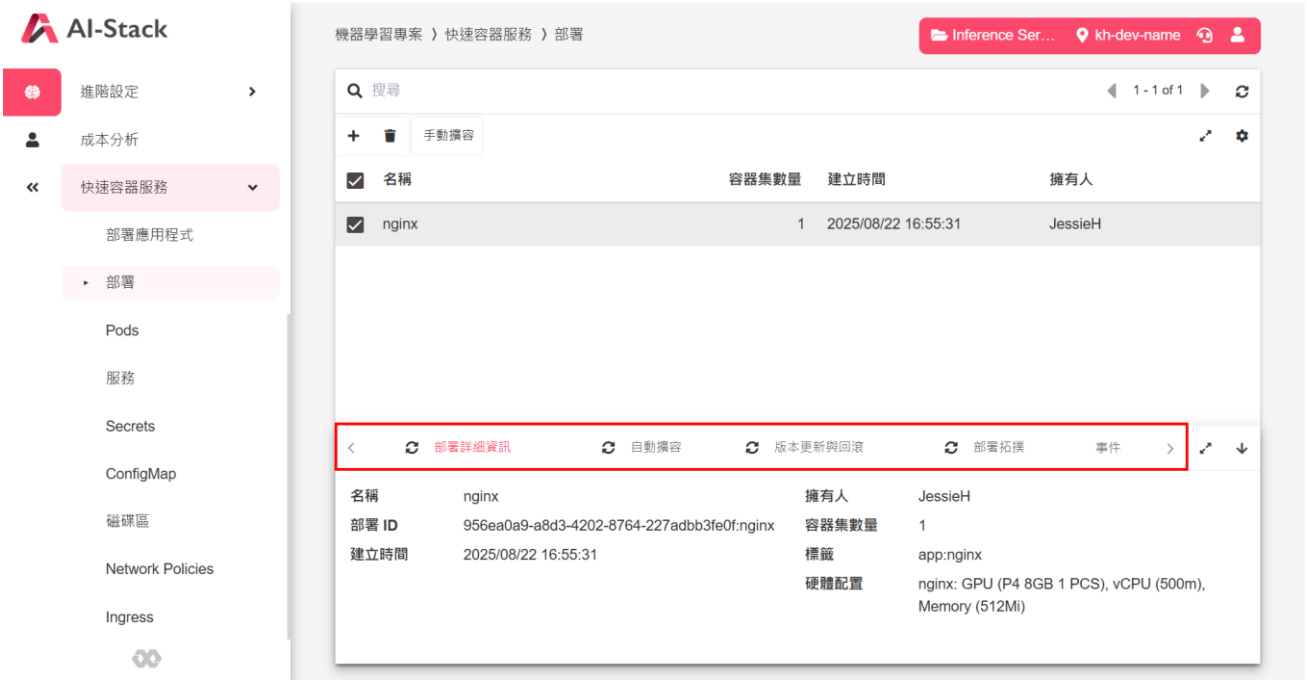
此處可查看當前所有的部署服務，管理 Pod 的副本數量和更新策略，確保應用程式按照期望狀態運作。

5.11.2.1 建立部署

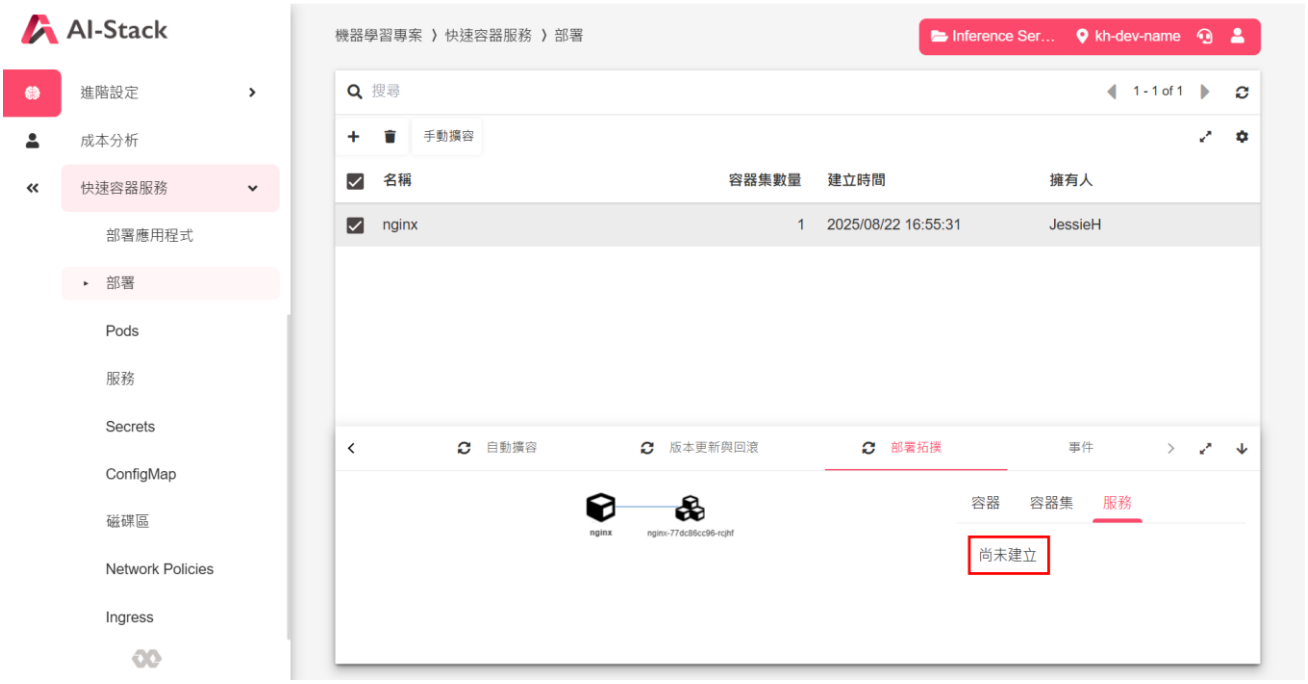
- 點選上方 ⊞ 會跳轉到部署應用程式。

5.11.2.2 管理部署

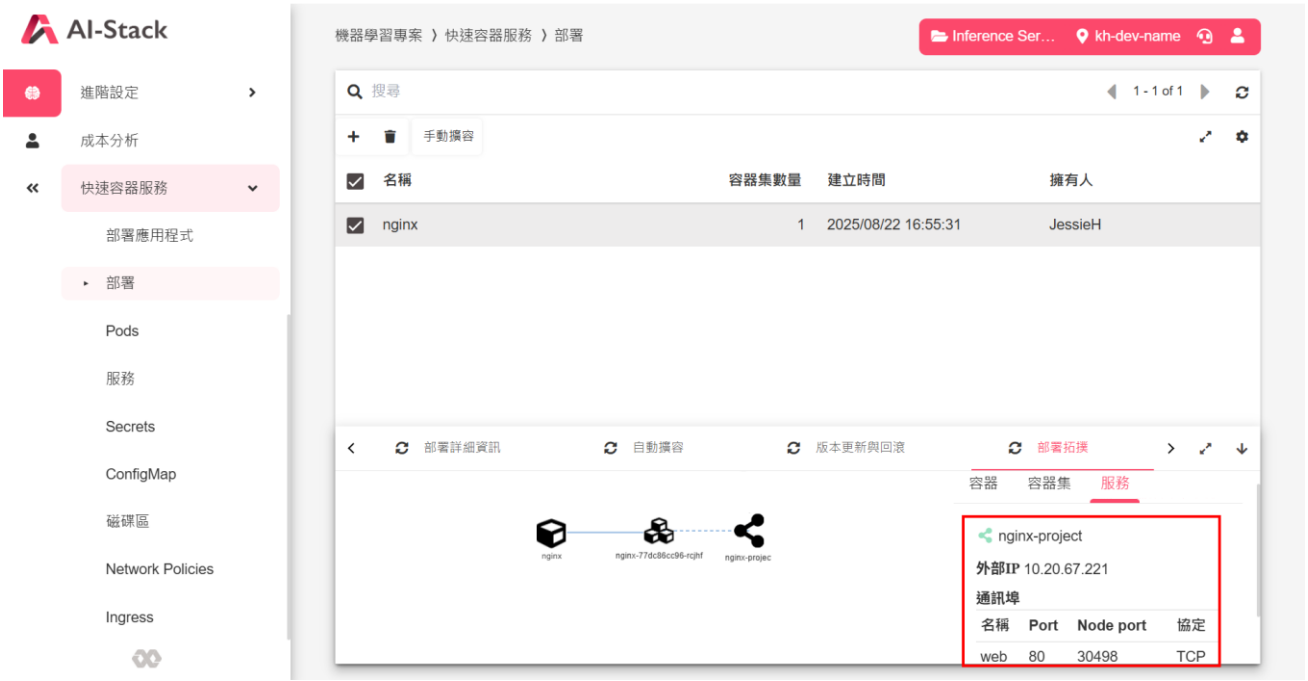
- 於清單中勾選目標部署，可於下方看到：
  - 部署詳細資訊
  - 自動擴容
  - 版本更新與回滾
  - 部署拓撲
  - 事件



- 若在 [部署拓撲] > [服務] 中看到「尚未建立」，接續建立服務進行建立。

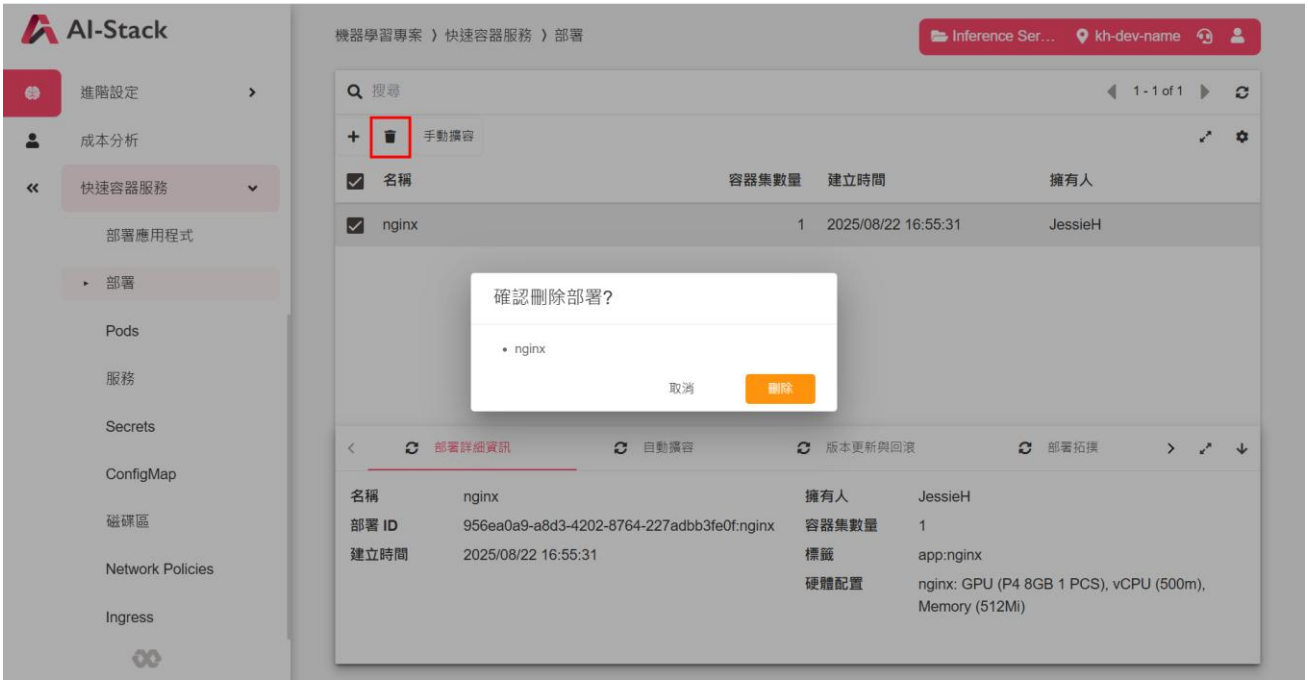


- 建立服務後在 [部署拓撲] 可看到拓撲圖上的圖示 ，點擊圖示可在右方檢視詳細資訊。
- 在 [服務] 的訊息中，可以知道透過 10.20.67.221:30498 可以存取 nginx 服務。



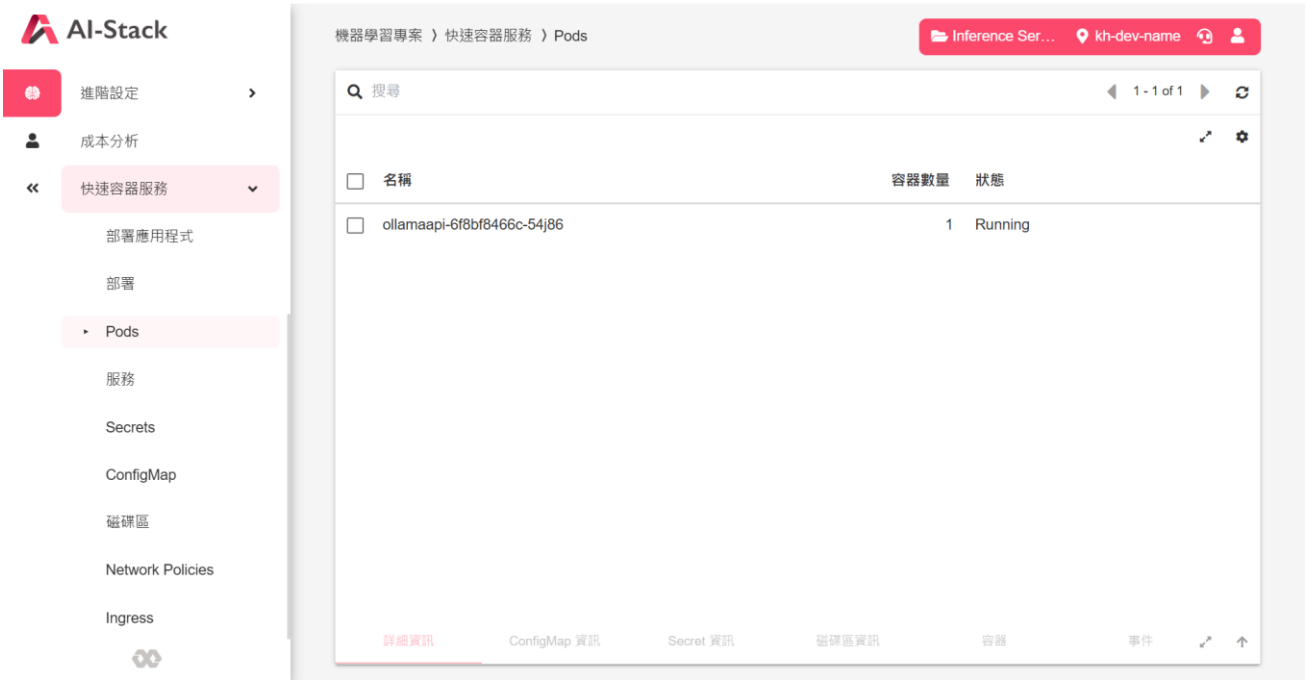
5.11.2.3 刪除部署

- 欲刪除部署時，可於清單中勾選目標部署，選定後點擊  將出現確認畫面，如下圖所示，確認為想要刪除的容器後再點擊 [刪除]。

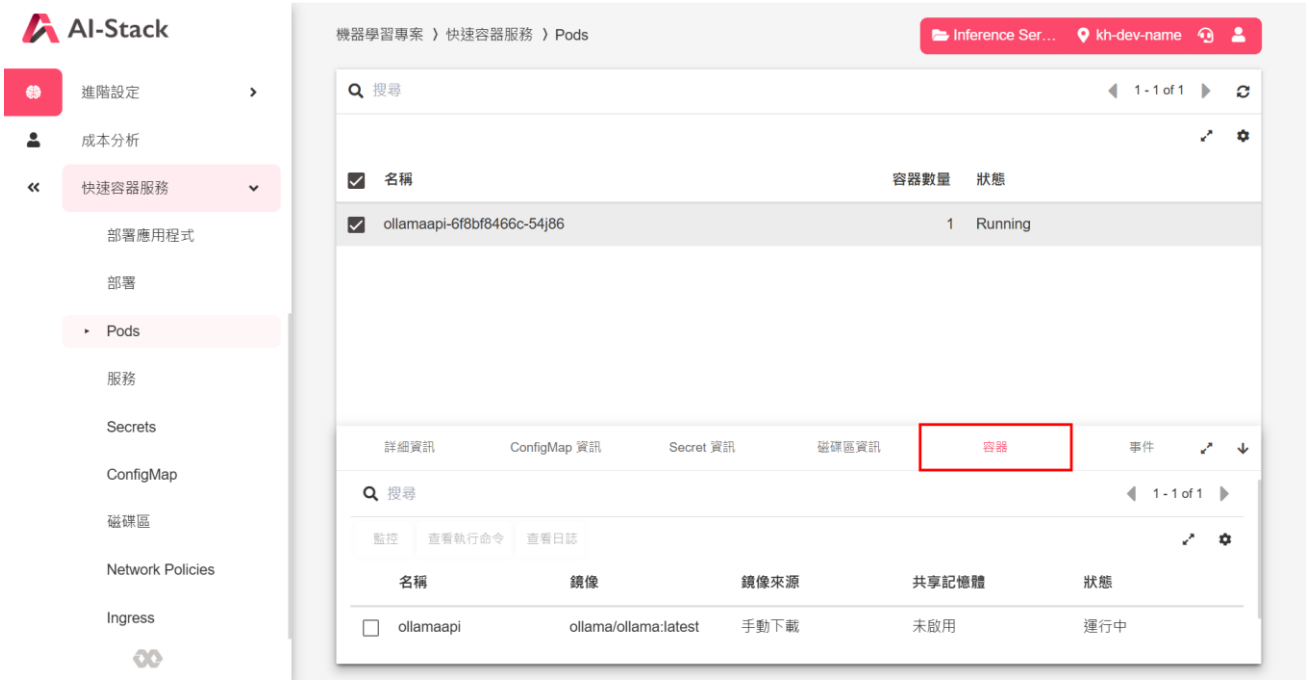


5.11.3 Pods

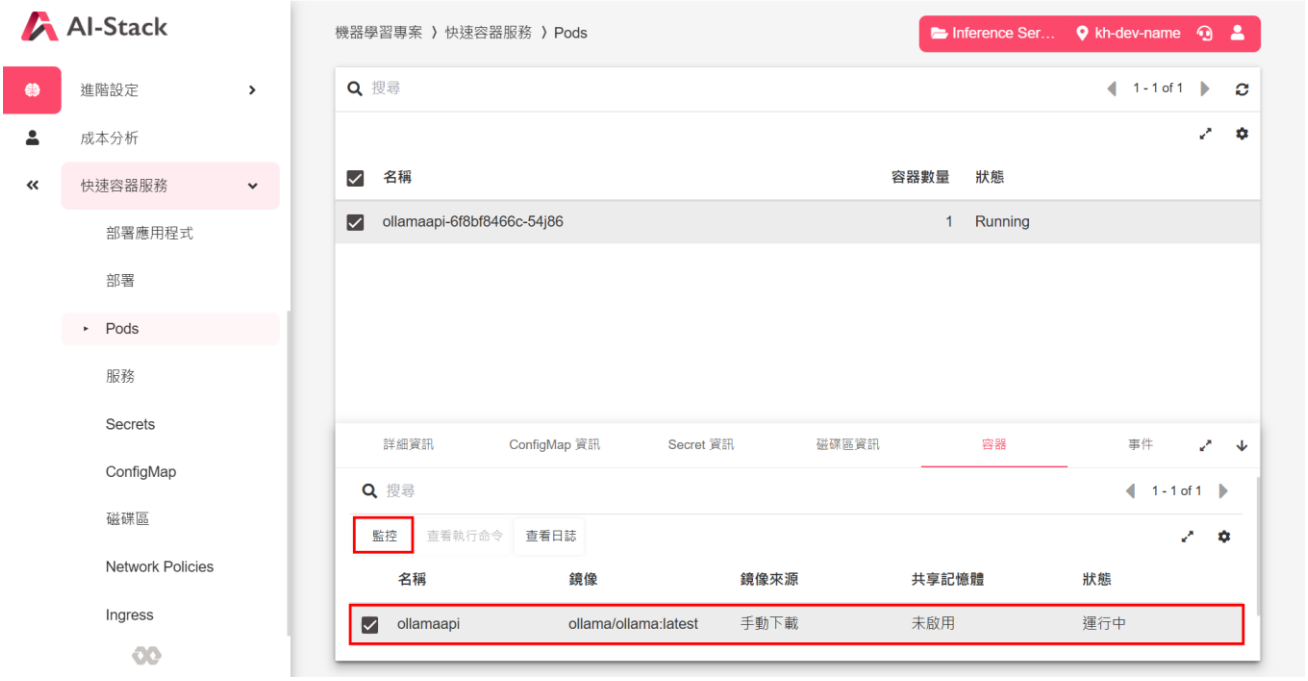
Pod 可運行一個或多個容器，並為每個容器提供共享的網路和儲存環境。此處可查看當前所有部署服務的容器集，及其包含的容器數量、狀態等相關資訊。

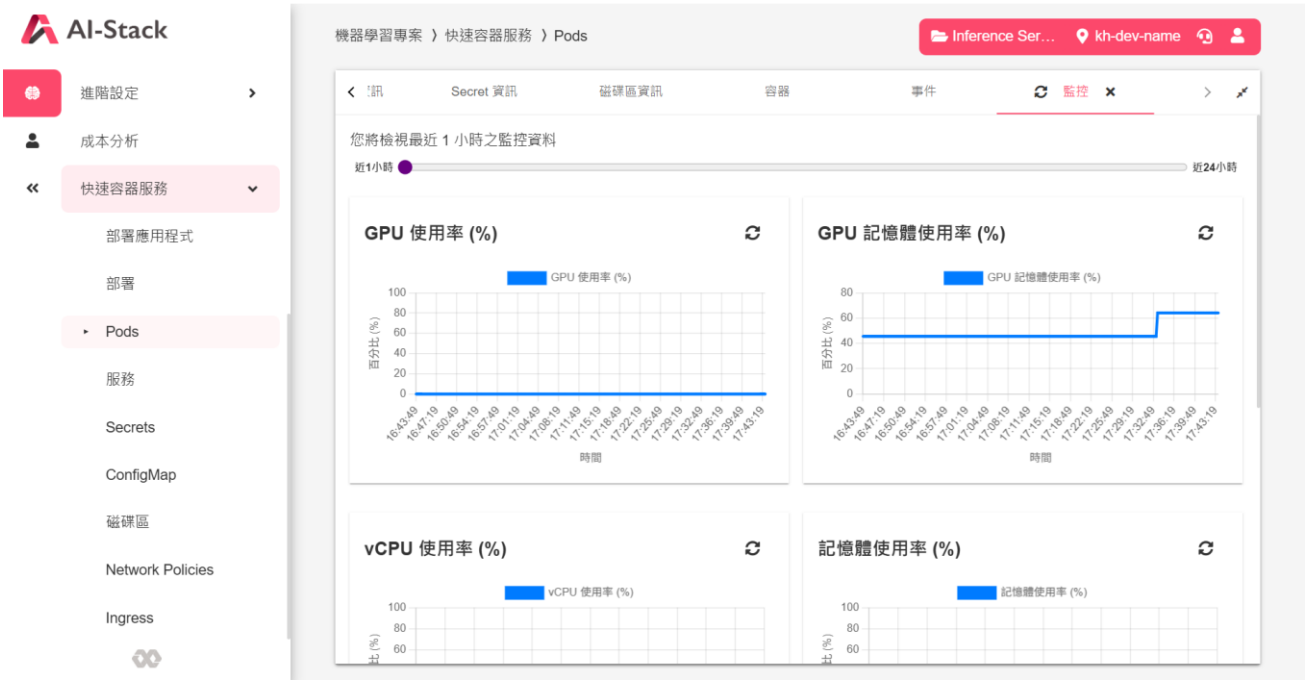


- 於清單中勾選目標 Pod，選定後點擊下方 [容器]。

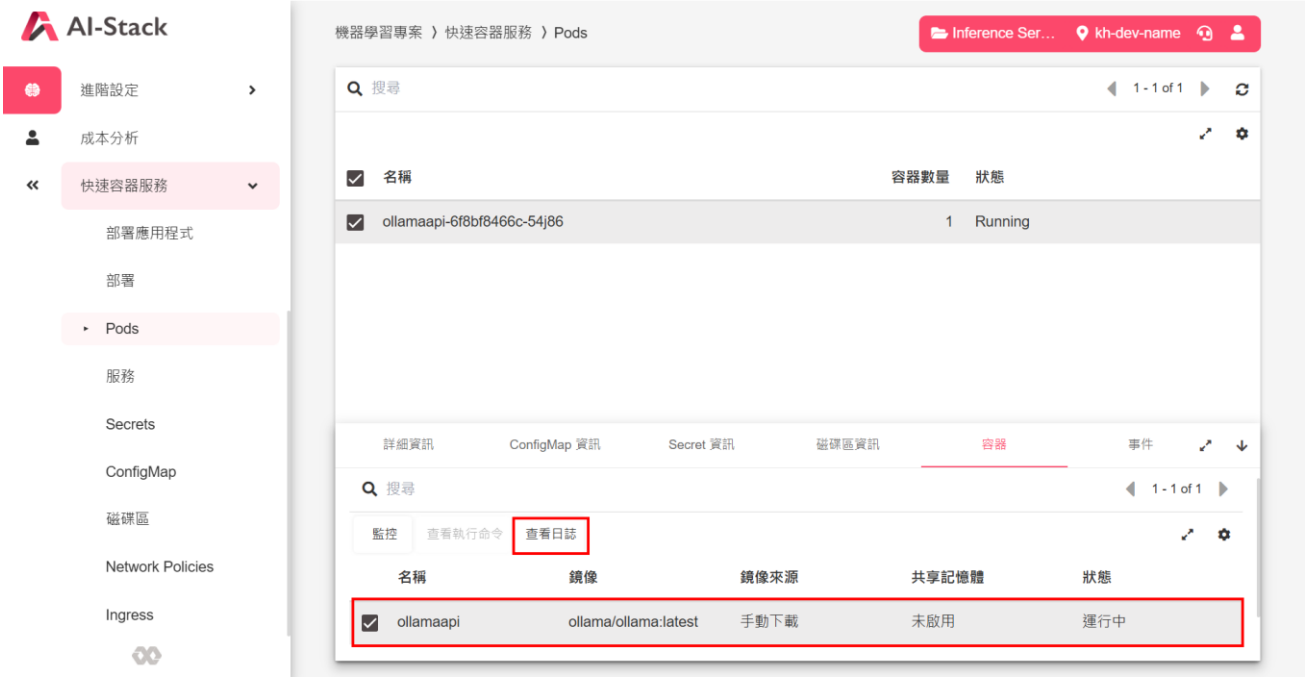


- 於清單中勾選目標容器，選定後點擊 [監控]，即可查看當前的各種資源使用率。





- 於清單中勾選目標容器，選定後點擊 [查看日誌]，即可顯示該容器的運行日誌。



5.11.4 服務

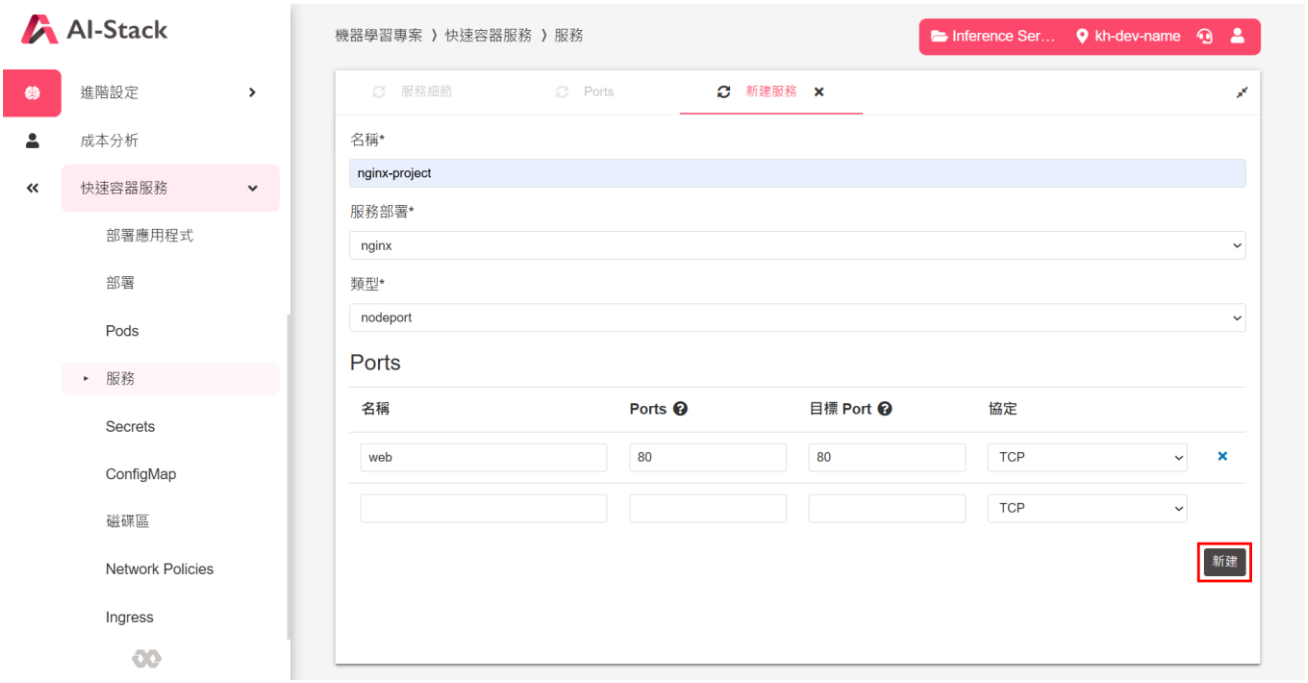
當 Pod 發生故障時，它們會被新的 Pod 取代，並且這些新 Pod 會有新的 IP 地址。擴展時會引入具有新 IP 地址的新 Pod，縮減則會移除 Pod。滾動更新也會用具有新 IP 的新 Pod 來取代現有的 Pod。此為 Kubernetes 的機制。此機制可能會產生大量的 IP 變動，因此需要 Service 來提供服務。

- Service 為 Pod 提供穩定的網絡連接。
- 每個 Service 都有自己的穩定 IP 地址、穩定的 DNS 名稱和穩定的端口。

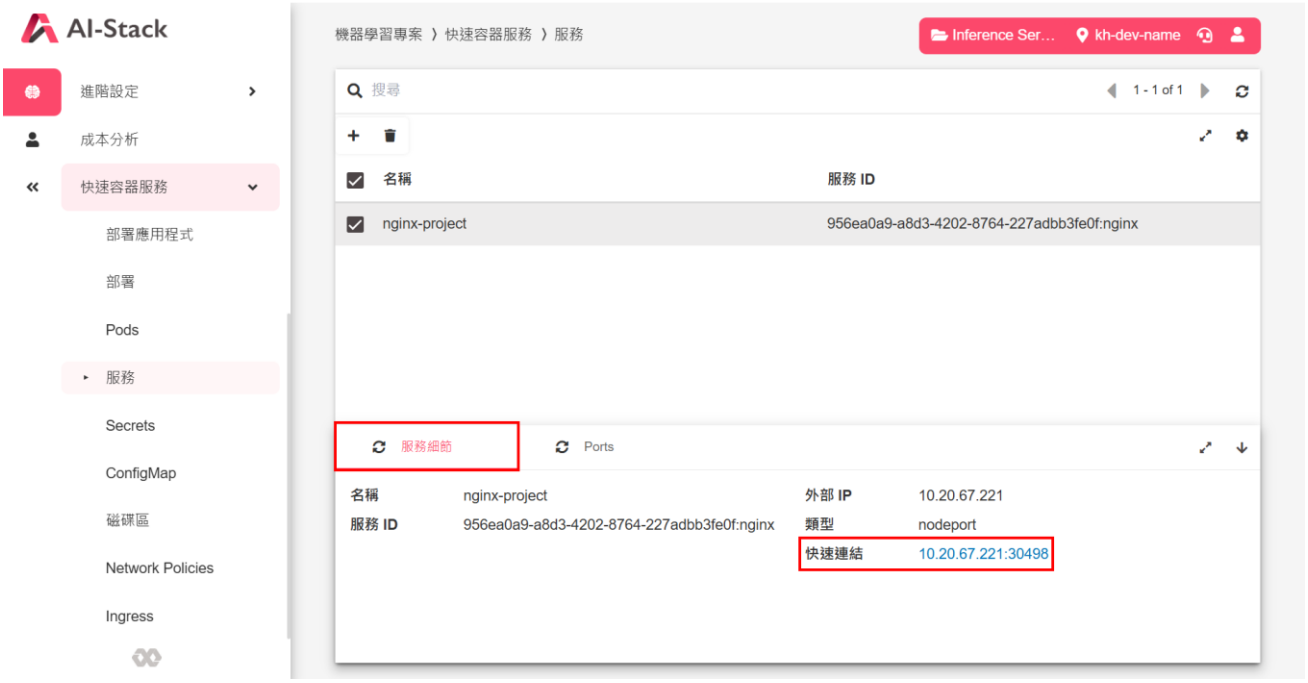
5.11.4.1 建立服務

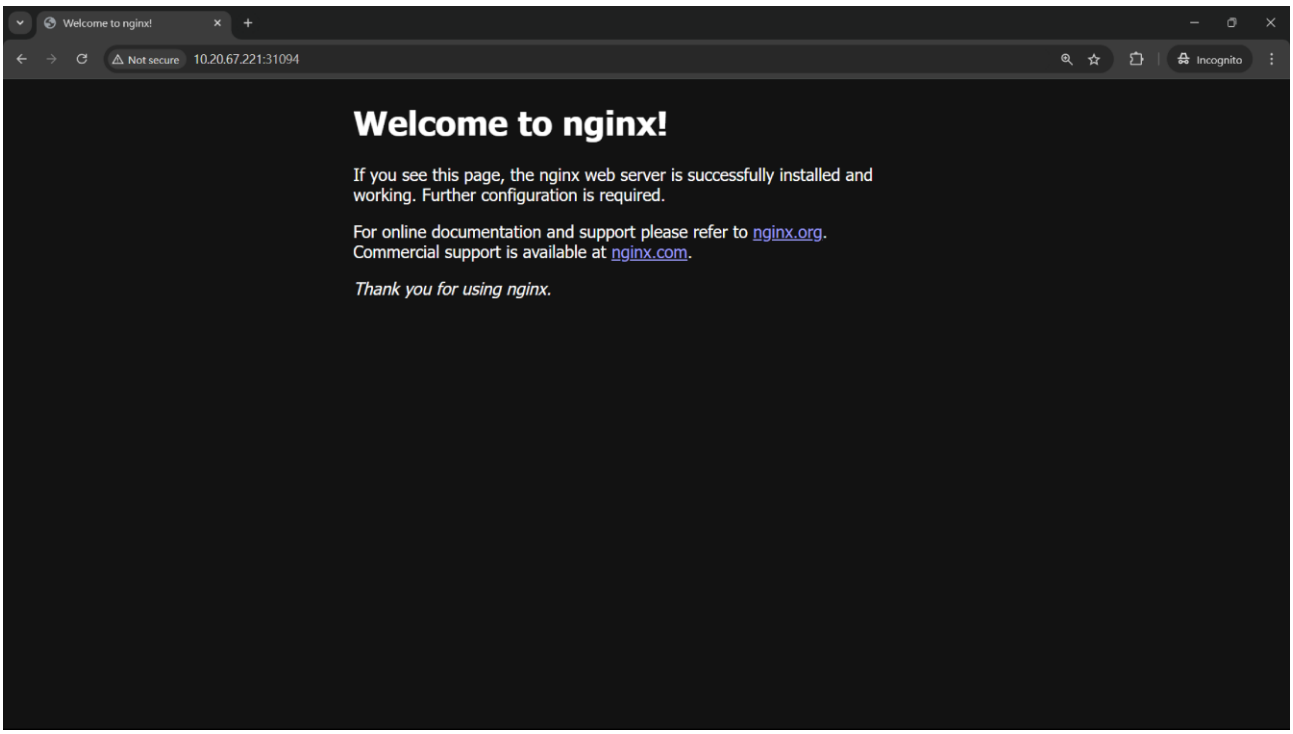
- 進入【服務】頁面點擊左上角 ⊞ 建立。

- 輸入方便識別的名稱（例：nginx-project）、選擇服務部署（例：nginx）。
- 選擇類型（此處範例選擇 nodeport）。
  - Nodeport：指定的連接埠暴露給外部，可以在叢集外部存取。
  - Clusterip：系統自動分配的虛擬 IP，只能在叢集內部存取。
- Port 名稱：可識別即可（例：web）。
- Port：要對外公開的通訊埠（例：80）。
- 目標 Port：對應服務（nginx container）暴露的通訊埠。
- 確認資料無誤後，點擊 [新建]。



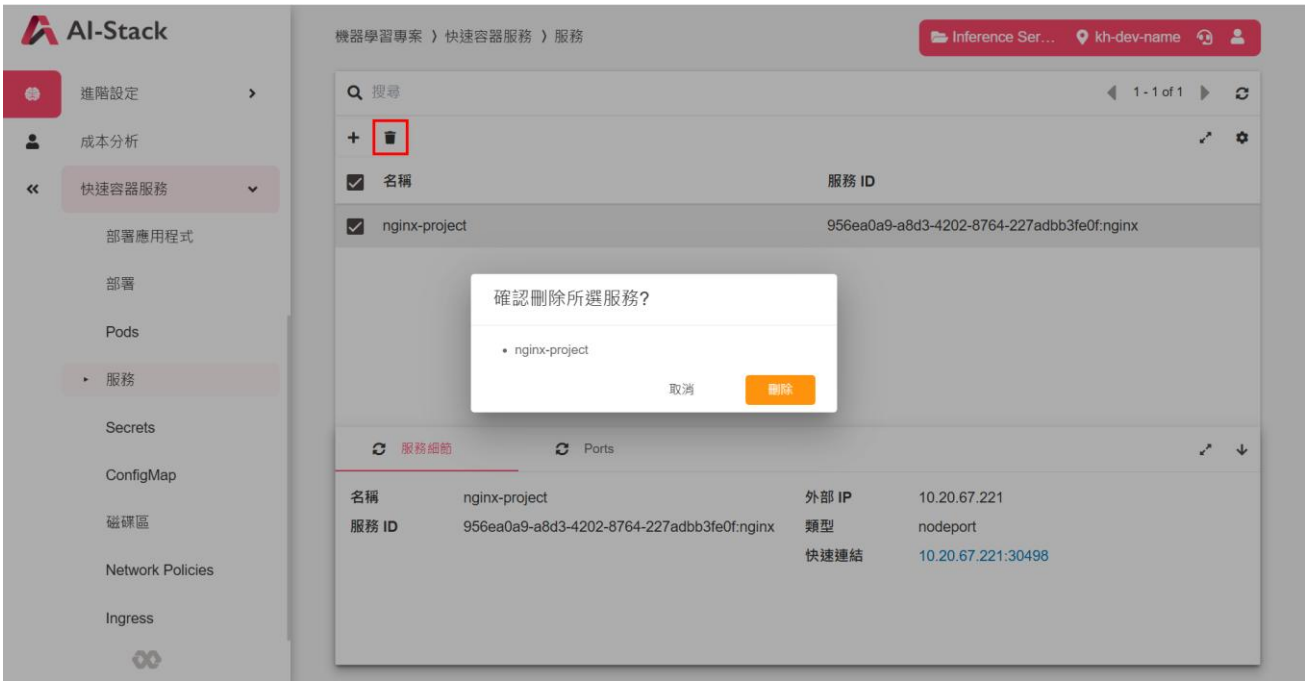
- 建立完成後，可選擇剛才建立的服務，即可在下方看到詳細資訊。
- 點選 [服務細節] > [快速連結] 可以在新分頁看到部署好的 nginx 畫面。





5.11.4.2 刪除服務

- 欲刪除服務時，可於清單中勾選目標服務，選定後點擊  將出現確認畫面，如下圖所示，確認為想要刪除的服務後再點擊 [刪除]。

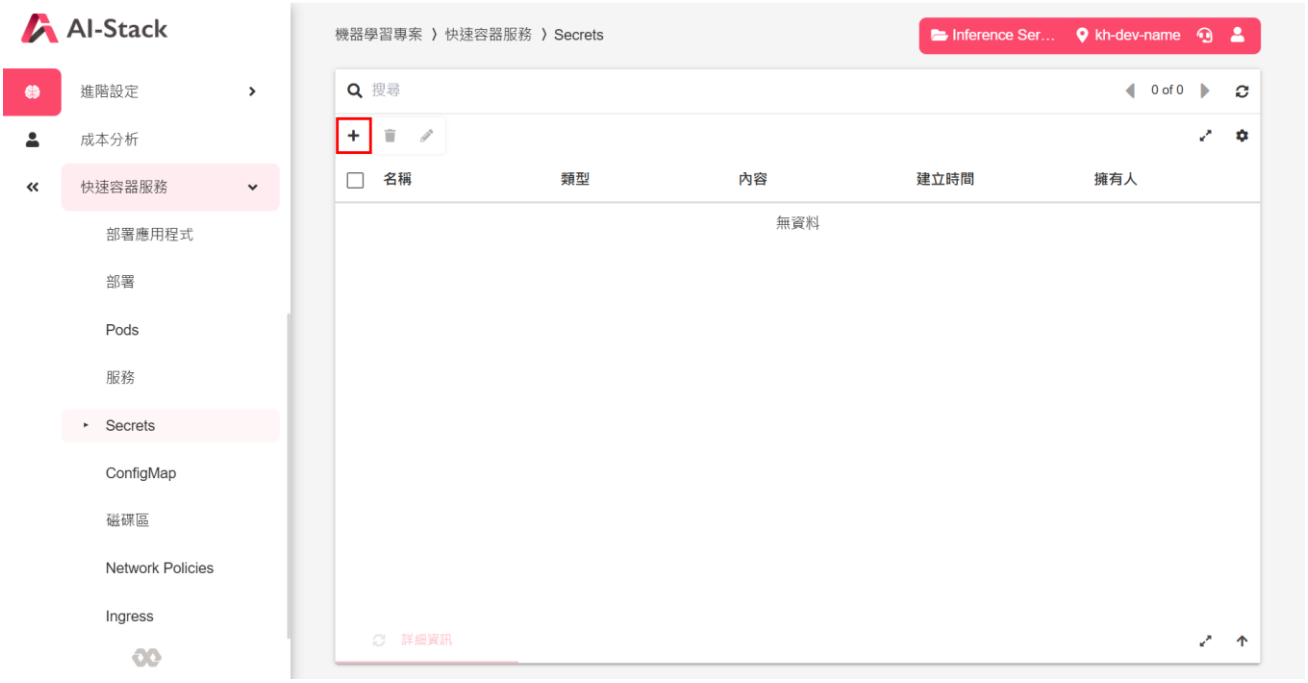


5.11.5 Secrets

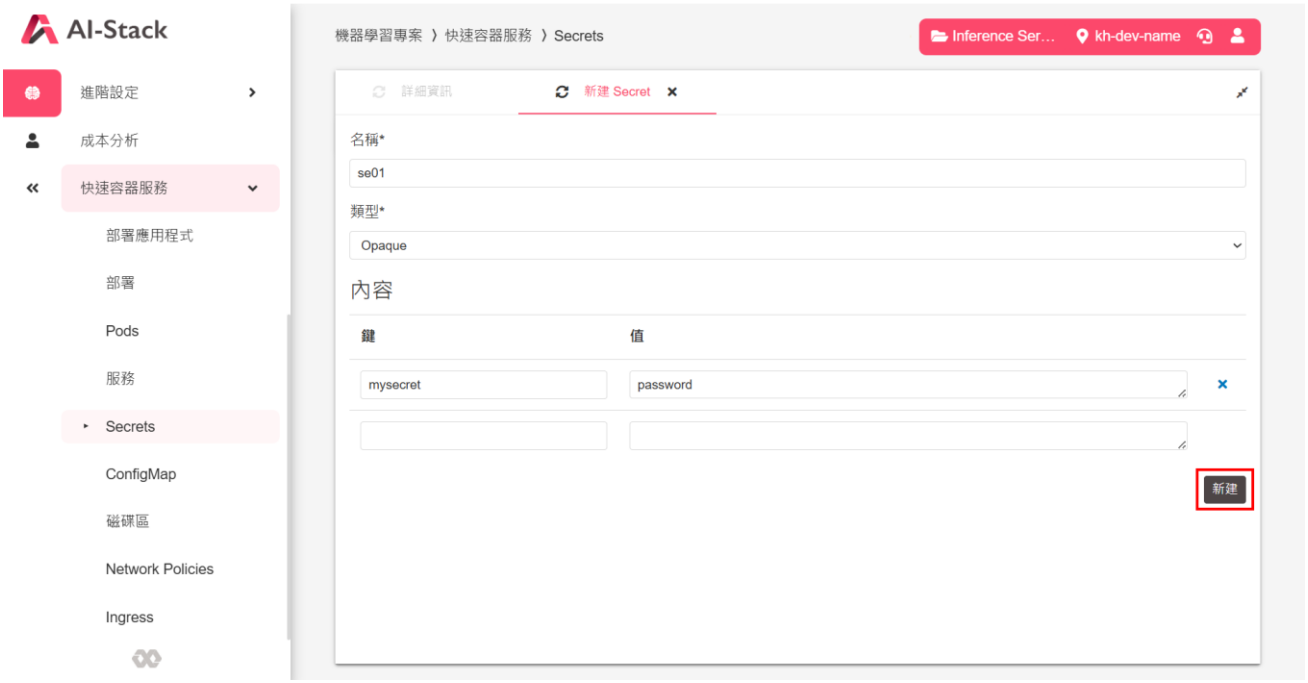
Secret 提供用於儲存**密碼、Token、金鑰等少量敏感資訊**。提供向 Pod 注入配置資訊的能力，而不是將它們以明文形式儲存在鏡像或設定檔中。

5.11.5.1 建立 Secret

- 進入【Secrets】頁面點擊左上角 ⊞ 建立。

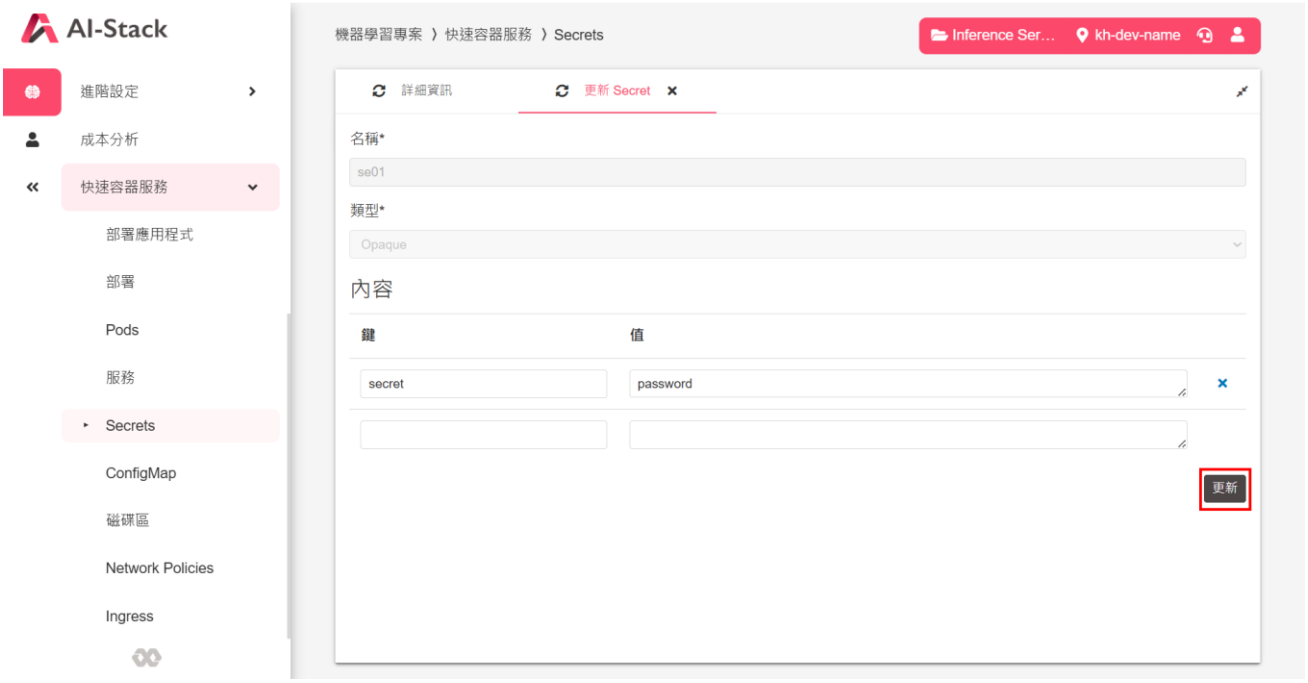


- 輸入方便識別的名稱。
- 選擇類型。
  - Opaque：用於儲存任意類型的鍵值對數據，如使用者名稱、密碼、API 金鑰等。
  - Docker Registry：用於存取私有 Docker 倉庫所需的認證訊息，包括使用者名稱、密碼等。
- 輸入鍵：輸入文件名。
- 輸入值：輸入文件內容。
- 確認資料無誤後，點擊 [新建]。



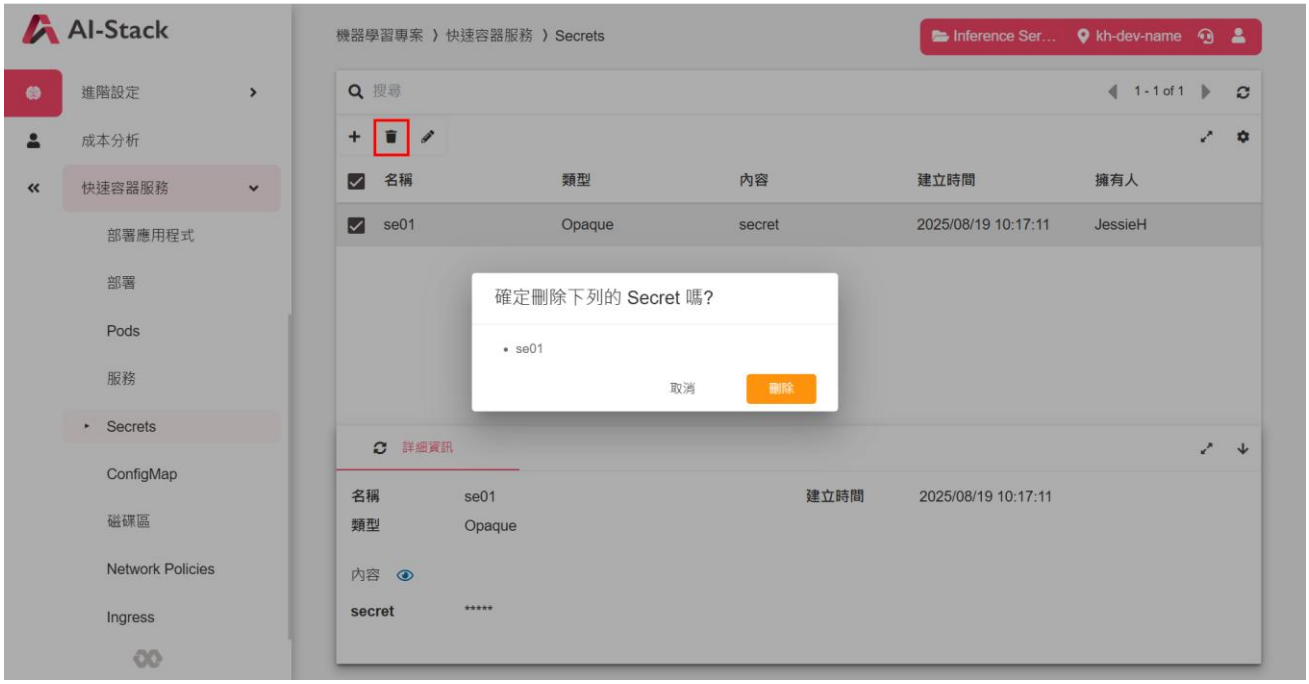
5.11.5.2 管理 Secret

- 於清單中勾選目標 Secret，可看到 [詳細資訊] 頁籤。
- 點擊  ，頁籤 [更新 Secret] 會出現，編輯鍵、值並確認資料無誤後，點擊 [更新]。



5.11.5.3 刪除 Secret

- 欲刪除 Secret 時，可於清單中勾選目標 Secret，選定後點擊  將出現確認畫面，如下圖所示，確認為想要刪除的 Secret 後再點擊 [刪除]。



5.11.6 ConfigMap

ConfigMap 是一個用於儲存配置參數的地方，這些配置參數可以在運行時無縫地注入到容器中，提供用來將**非機密性**的鍵值對、文字檔案或以特定格式組織的設定文件儲存在 Pod 之外，如環境變數、命令列參數、服務端口、主機名、帳戶名稱等，幫助使用者管理應用程式的設定資訊，可以在運行時存取 ConfigMap 中的配置參數，以便動態調整應用程式的行為。

5.11.6.1 建立 ConfigMap

- 進入【ConfigMap】頁面點擊左上角 ⊞ 建立。

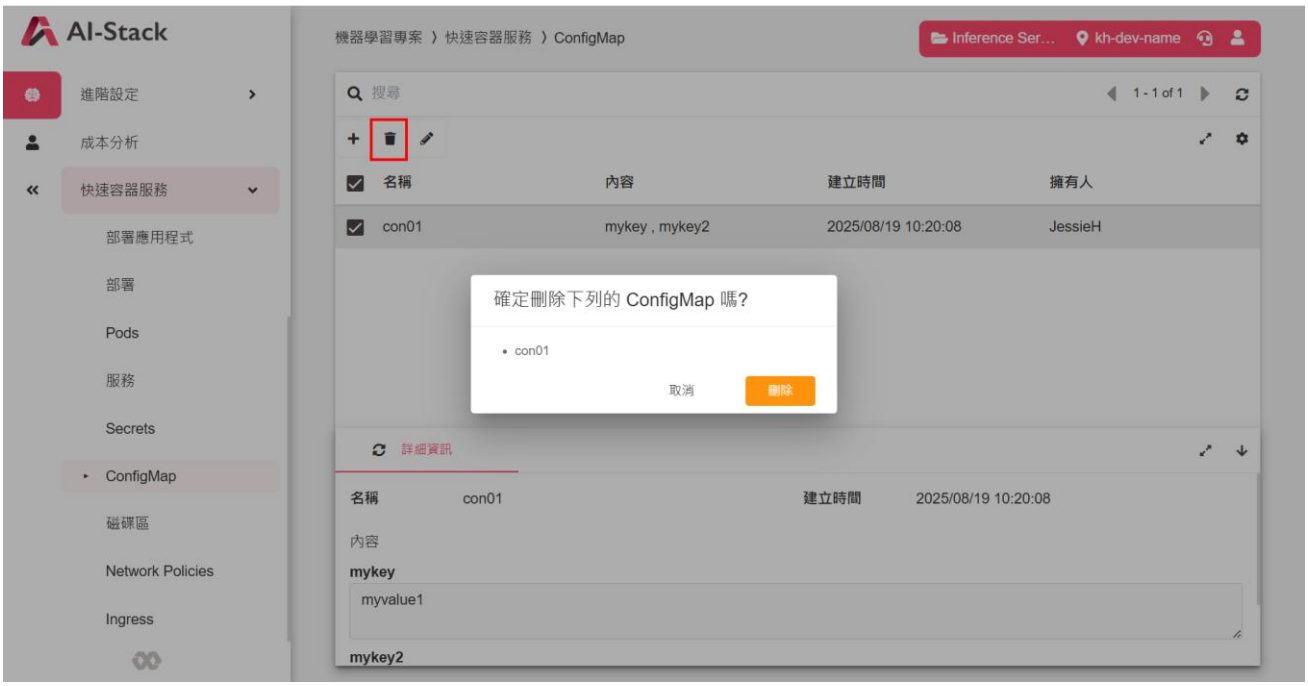
- 輸入方便識別的名稱。
- 選擇類型。
- 輸入鍵，可以由大小寫英文字母、數字、半形符號「-」、「.」和「_」組成的任意名稱。
- 輸入值，可以包含任何內容。
- 確認資料無誤後，點擊 [新建]。

5.11.6.2 管理 ConfigMap

- 於清單中勾選目標 ConfigMap，可看到 [詳細資訊] 頁籤。
- 點擊  ，頁籤 [更新 ConfigMap] 會出現，編輯鍵、值並確認資料無誤後，點擊 [更新]。

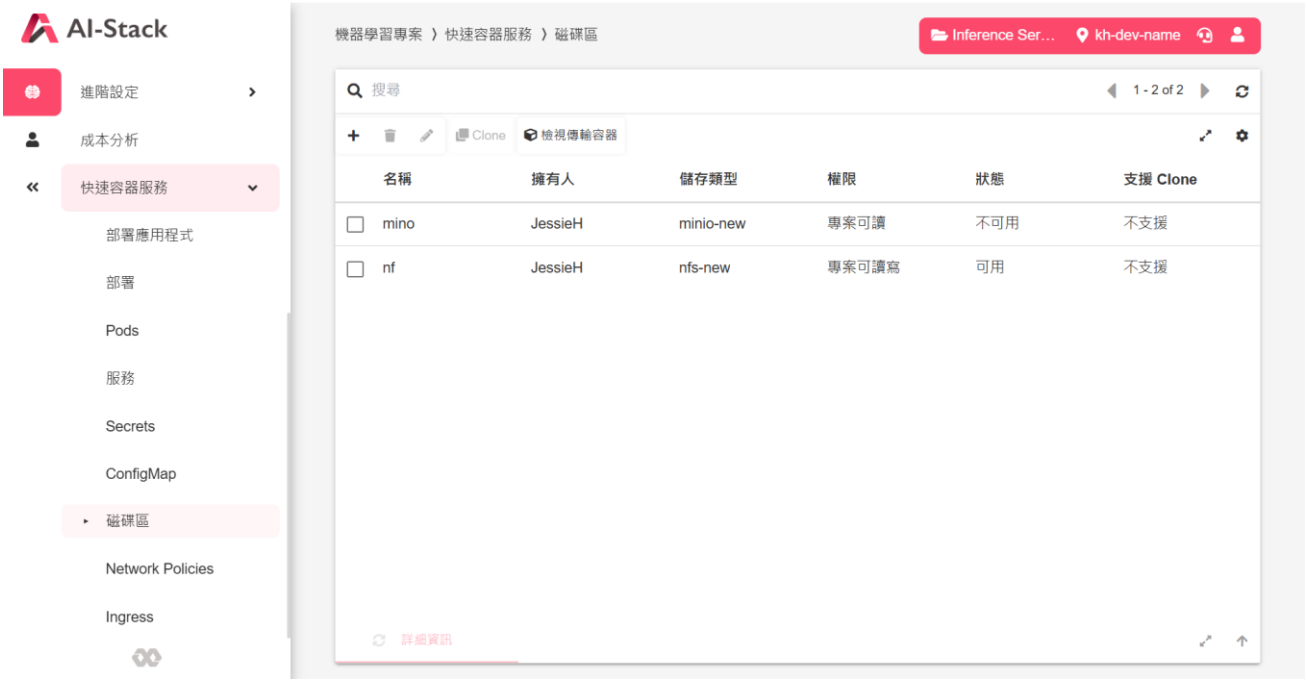
5.11.6.3 刪除 ConfigMap

- 欲刪除 ConfigMap 時，可於清單中勾選目標 ConfigMap，選定後點擊  將出現確認畫面，如下圖所示，確認為想要刪除的 ConfigMap 後再點擊 [刪除]。



5.11.7 磁碟區

磁碟區（Volume）可以實現容器生命週期之外持久保存資料、決定掛載路徑、建立傳輸容器等，詳細操作步驟參考儲存管理。

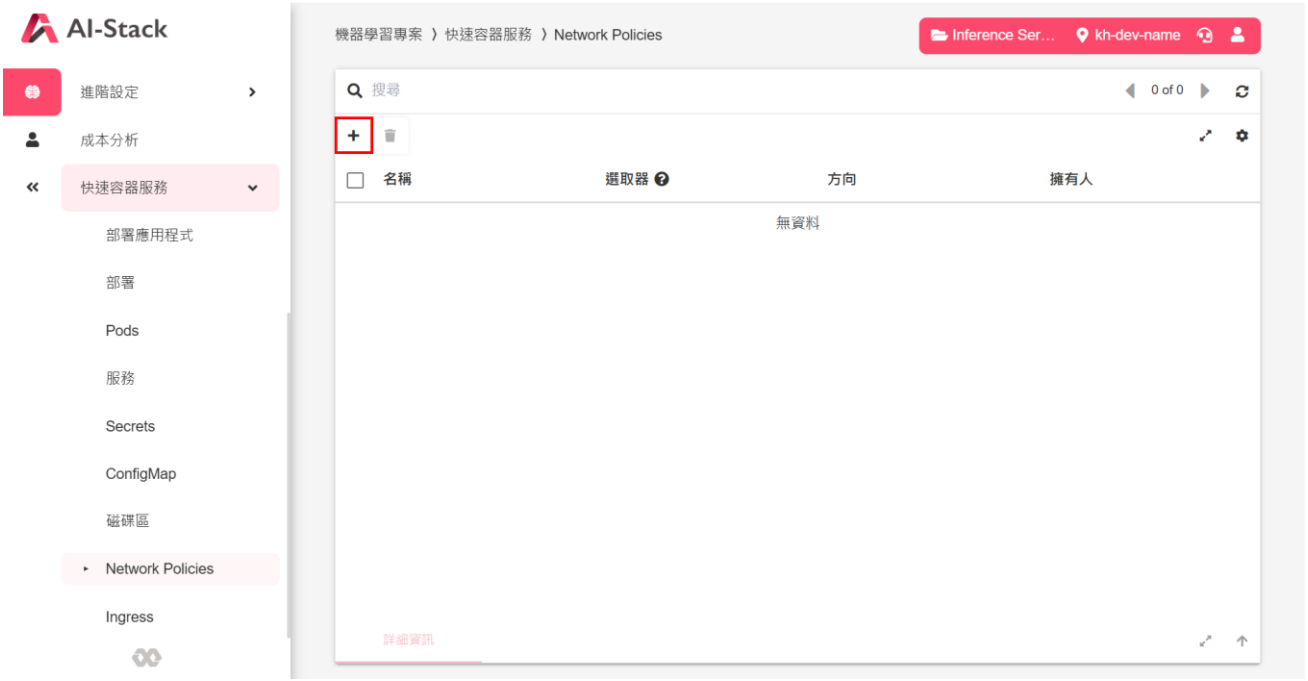


5.11.8 NetworkPolicies

預設情況下，所有 Pod 是非隔離的，即任何來源的網路流量都能夠存取 Pod，沒有任何限制。NetworkPolicy 是一種關於 Pod 間及與其他網路端點間所允許的通訊規則的規範，允許使用者定義一組規則，規定選定 Pod 所允許的通訊，從而實現網路隔離和安全性。

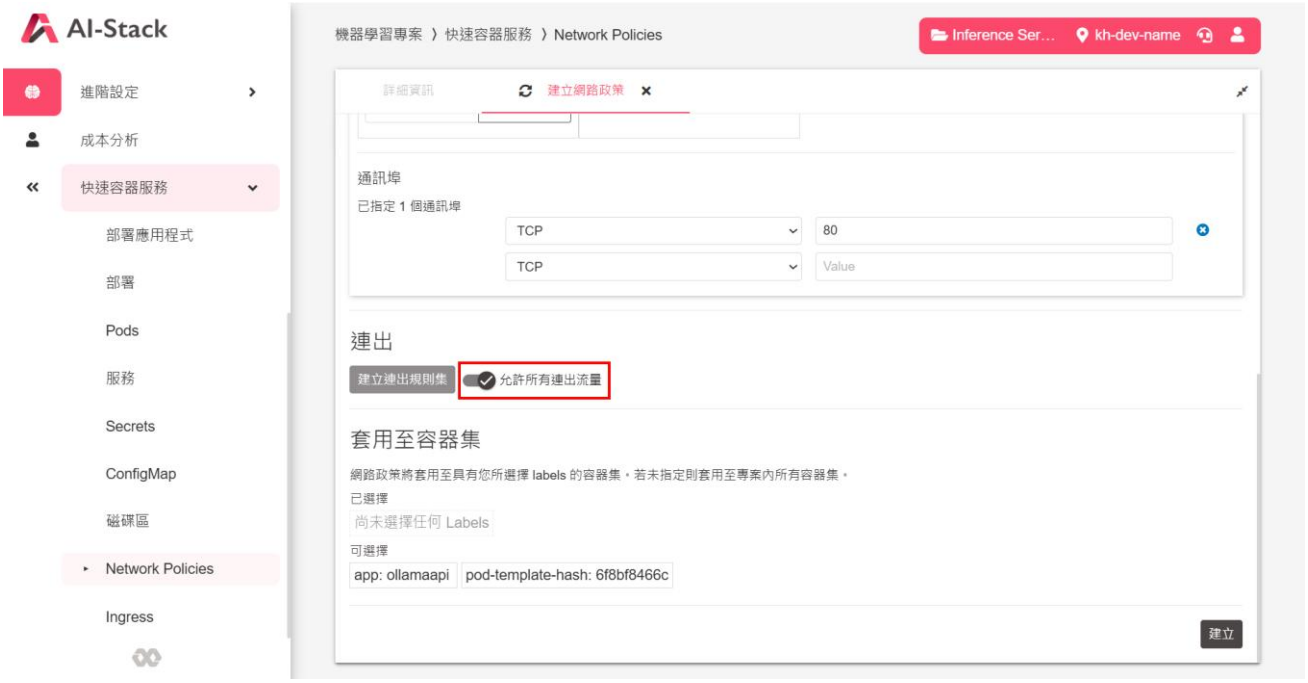
5.11.8.1 建立 NetworkPolicy

- 進入【NetworkPolicies】頁面點擊左上角 ⊞ 建立。

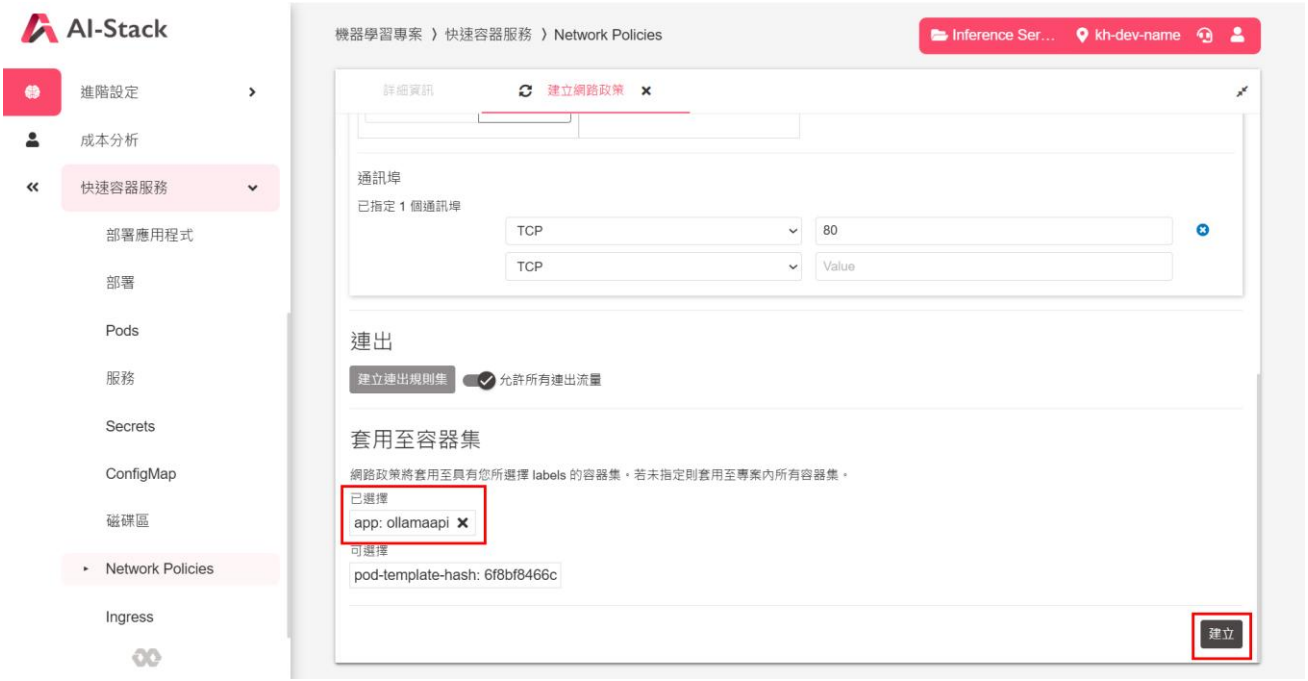


- 輸入方便識別的名稱。
- 建立連入、連出規則。
  - 套用網段：選定 IP CIDR 範圍以用作連入、連出的流量來源。
  - 排除網段：在連入、連出的範圍中排除指定的網段。
  - 通訊埠：允許 Pod 連接到規定的通訊埠。

- 若連入、連出不需要設置，可以直接選擇 [允許所有連入流量] 或 [允許所有連出流量]。

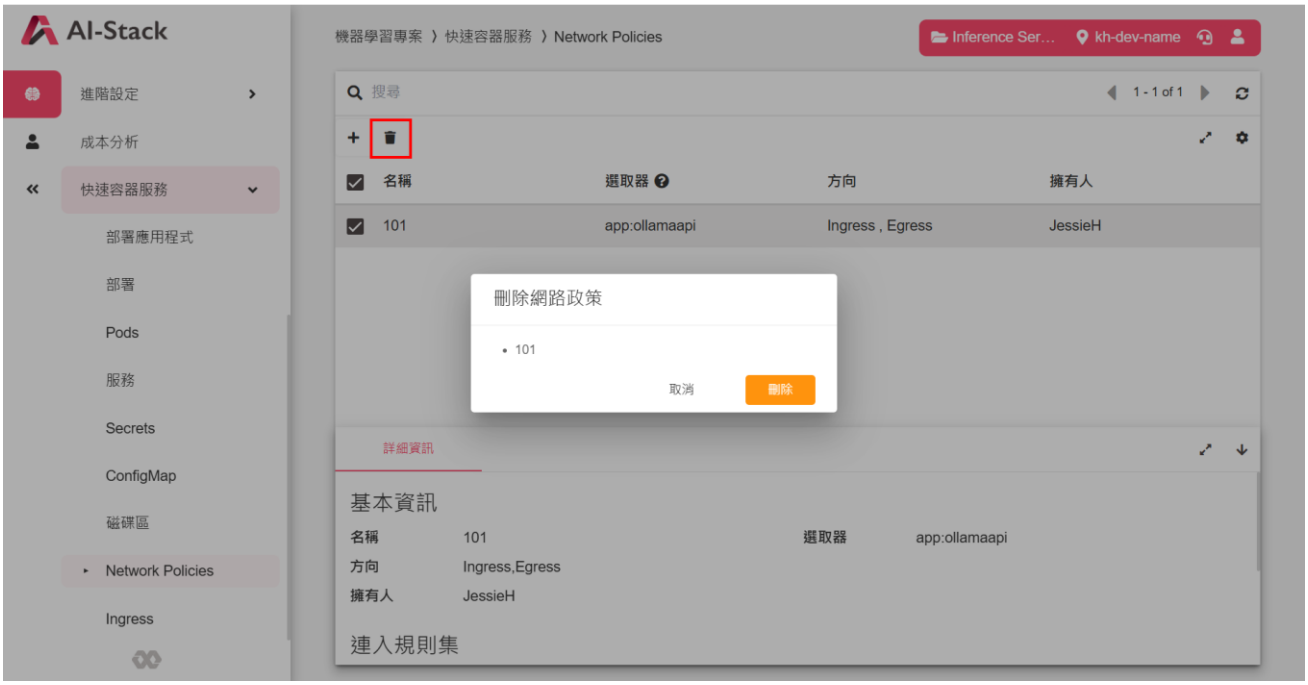


- 選擇套用的容器集，若沒有選擇會套用到所有容器集。
- 確認資料無誤後，點擊 [建立]。



5.11.8.2 刪除 NetworkPolicy

- 欲刪除 NetworkPolicy 時，可於清單中勾選目標 NetworkPolicy，選定後點擊  將出現確認畫面，如下圖所示，確認為想要刪除的 NetworkPolicy 後再點擊 [刪除]。



5.11.9 Ingress

Ingress 提供從叢集外部到叢集內服務的 HTTP 和 HTTPS 路由，可以將服務對應到某個 Domain，讓使用者以 URL 存取服務。

5.11.9.1 建立 Ingress

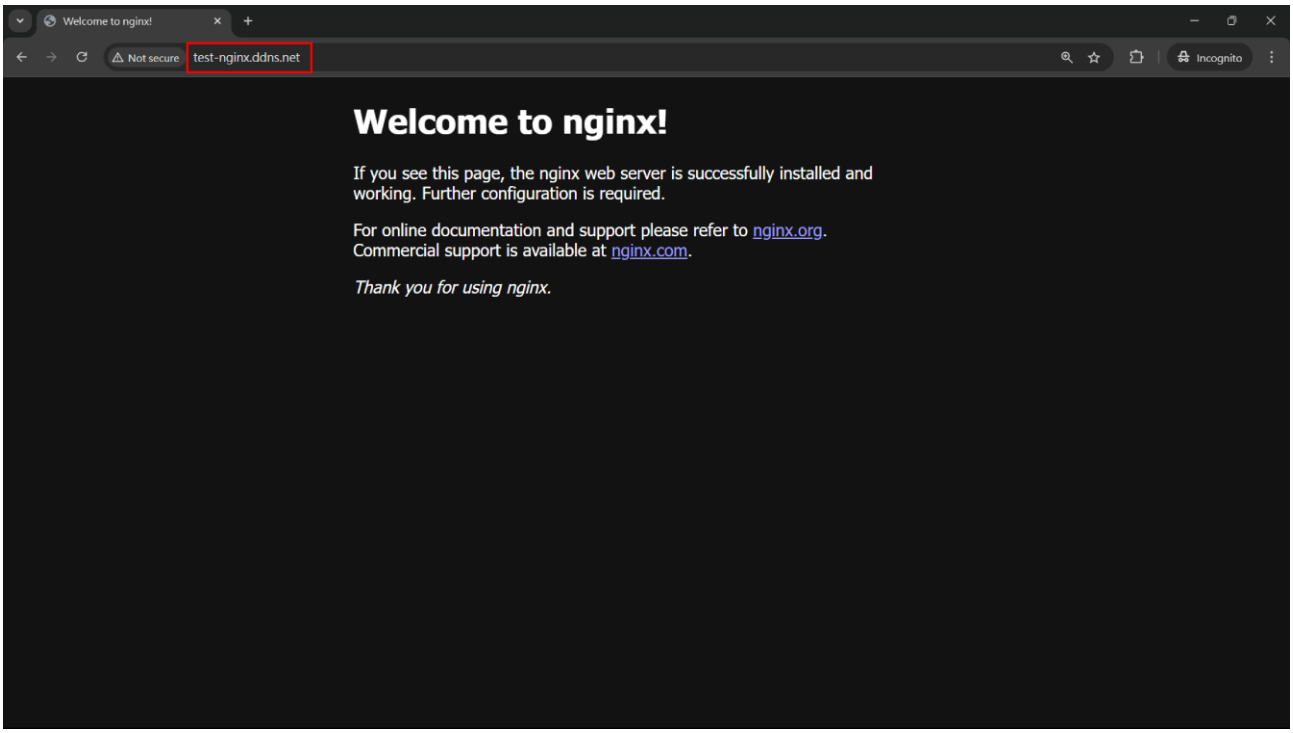
- 進入【Ingress】頁面點擊左上角 ⊞ 建立。

- 輸入方便識別的名稱。
- 路徑重寫：指根據配置規則修改或替換請求的路徑的過程。

- 設定網域，填入準備好的網域。

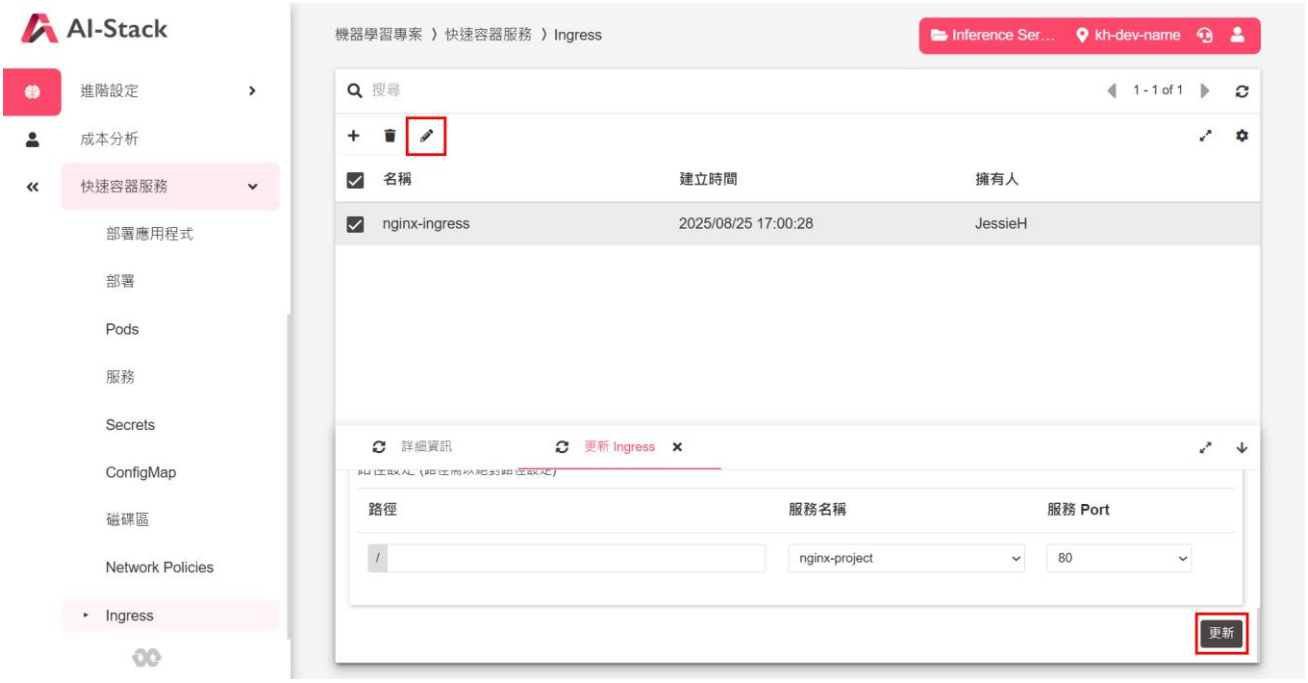
- 設定路徑，選擇需要建立的服務。
- 確認資料無誤後，點擊 [新建]。

- 建立完成後，可以透過瀏覽器，以網域存取服務。



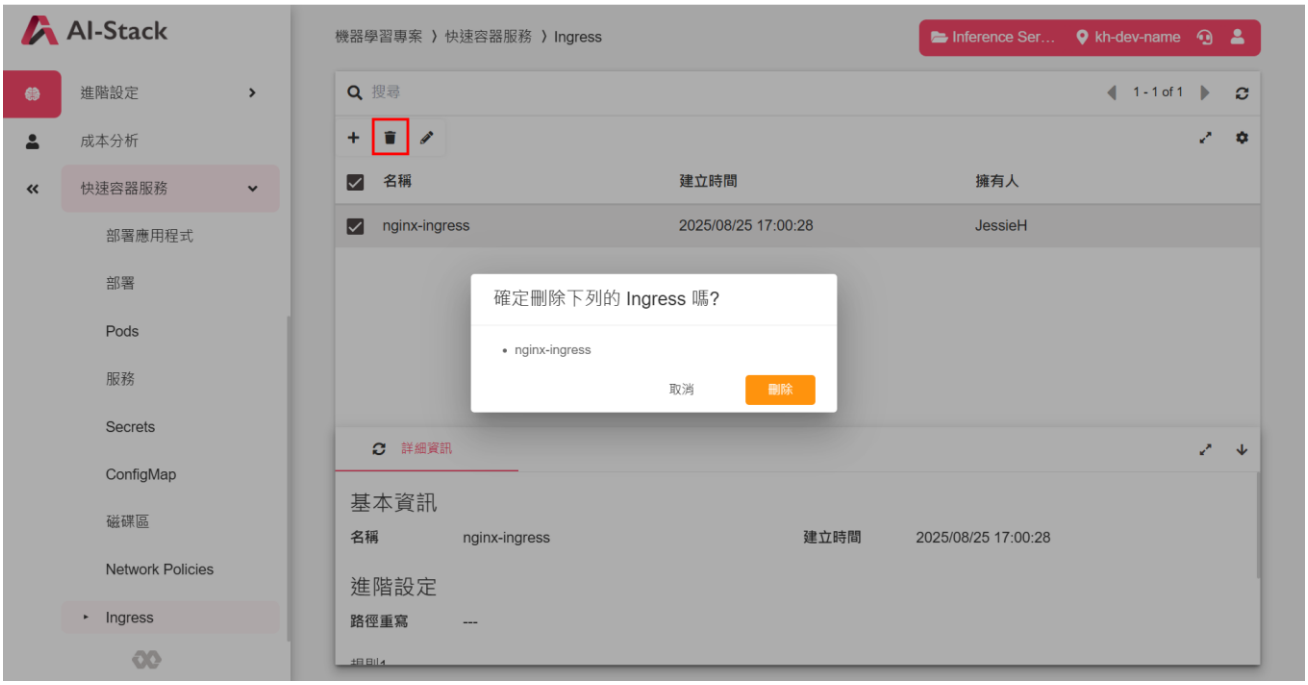
5.11.9.2 管理 Ingress

- 於清單中勾選目標 Ingress，可看到 [詳細資訊] 頁籤。
- 點擊  ，頁籤 [更新 Ingress] 會出現，編輯並確認更新內容無誤後，點擊 [更新]。



5.11.9.3 刪除 Ingress

- 欲刪除 Ingress 時，可於清單中勾選目標 Ingress，選定後點擊  將出現確認畫面，如下圖所示，確認為想要刪除的 Ingress 後再點擊 [刪除]。



6. 分佈式訓練叢集

分佈式訓練已成為深度學習領域因應大規模模型與數據運算需求的核心技術。AI-Stack 自 4.26.0 版本起，正式推出「分佈式訓練叢集」功能（Enterprise 版本專屬），旨在協助開發者提升模型訓練的靈活性與效率。此模組能夠將訓練工作負載分散至多個容器或節點，有效縮短訓練時間，並突破單一設備的資源限制。

主要功能與特性

- 統一環境掛載

  一鍵掛載使用者 Home 目錄，讓雲端容器可直接存取本地開發程式碼、資料集與腳本，省略資料同步步驟，提升雲端與本地工作流程的一致性。

- 支援主流訓練框架

  原生支援 Horovod 和 DeepSpeed 等主流分佈式訓練框架，簡化多機協同訓練任務的設定流程。未來將擴充支援更多 AI / HPC 主流訓練框架，以因應多樣化的大模型訓練需求。

- 動態資源擴縮

  容器叢集資源可依任務需求彈性調整，支援即時擴增或縮減容器數量，無需手動設定 Hostfile，有效提升資源使用率並降低操作複雜度。

- 操作介面與監控

  提供直觀的 UI 介面協助用戶建立、管理和監控訓練叢集。內建資源監控與效能分析工具，有助於訓練過程中的問題診斷、效能優化與成本控管。

「分佈式訓練叢集」模組設計目標為降低大規模分佈式訓練的使用門檻，提升開發效率，協助團隊將重心集中於模型創新與優化。

* 備註：自 4.23.0 版本起提供的分佈式訓練 Alpha 版本已於 4.26.0 版停止提供，建議用戶全面升級至新版正式功能。

6.1 建立分佈式訓練叢集

開始建立分佈式訓練容器叢集前，須確認以下內容，若需進行相關設定，請洽管理者，並建議參考《AI-Stack Enterprise 機器學習協作管理平台 管理者操作手冊》中相關章節。

- 確認專案內具備可用的 [MLS 規格]，且規格內不可選擇共享模式的 GPU。建議參考章節：
  - 第 8.1.1 章〈MLS 規格〉: 如何設定建立容器時可選用的 MLS 規格 (包含 GPU 型號、數量、CPU 核心數、記憶體等)。
  - 第 4.1 章〈專案列表〉: 如何設定專案可用資源。請確保該規格可用於您的專案。
- 確認【公用鏡像列表】中備有適合 DeepSpeed、Horovod 框架運行的鏡像。建議參考章節：
  - 第 6 章〈鏡像管理〉: 如何操作 [鏡像匯入工具] 匯入鏡像，並透過【MLS 樣板】設定將鏡像上架使用者公用鏡像列表。

  若無，亦可先使用容器服務建立容器並自行安裝訓練框架後，建立自定義鏡像，並於建立叢集時選用已安裝訓練框架的自定義鏡像。

* 注意：

1. 單次創建的容器資源規格需保持一致，不同批次可配置不同規格，但建議同一叢集中的容器規格一致，以確保最佳性能與穩定性。
2. 分佈式訓練適用的容器資源規格限制 GPU 以片為最小單位，不支援部署訓練容器於 GPU 共享模式節點中。

6.1.1 建立容器叢集

- 進入【容器管理】>【分佈式訓練叢集】，點按 [建立叢集] 建立容器叢集。

① 基本設定

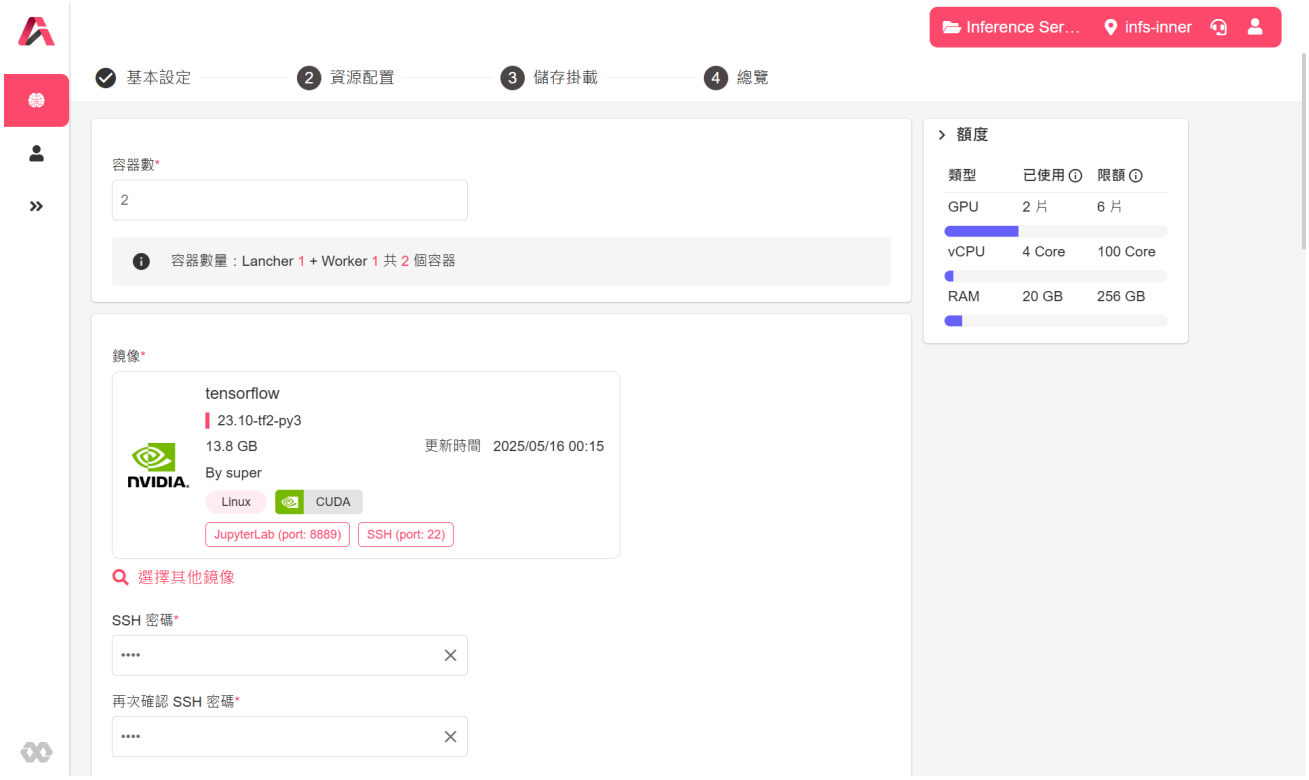
- 選擇要建立的**叢集類型**，目前支援 Horovod、DeepSpeed 框架。
- 輸入**叢集名稱**，只能使用小寫英數字元及 - 符號，且開頭和結尾必須是英數字元。
- 輸入**描述**（選填）。
- 設定**排程**時間。
  - **起始時間**：預設立即執行，亦可指定特定日期及時間。
  - **結束時間**：預設不指定結束時間，亦可指定結束日期、時間或指定結束時數。
- 點擊 [下一步] 設定資源配置。

② 資源配置

- 設定**容器數**。容器數量必須大於等於 2，包含一個 launcher 容器及其餘 worker 容器。

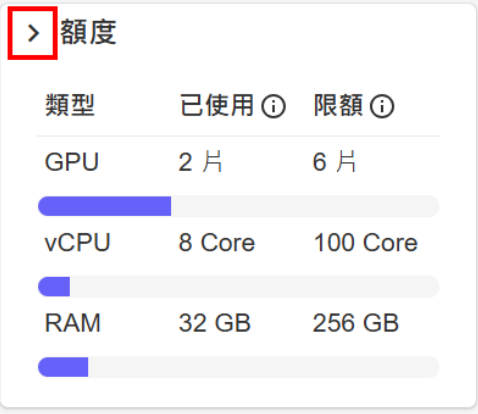
Launcher 容器負責統一設定容器啟動參數，執行啟動指令及控制通訊流程；而 worker 容器則負責執行實際的訓練。設定 launcher 及 worker 容器可確保整體訓練流程穩定。

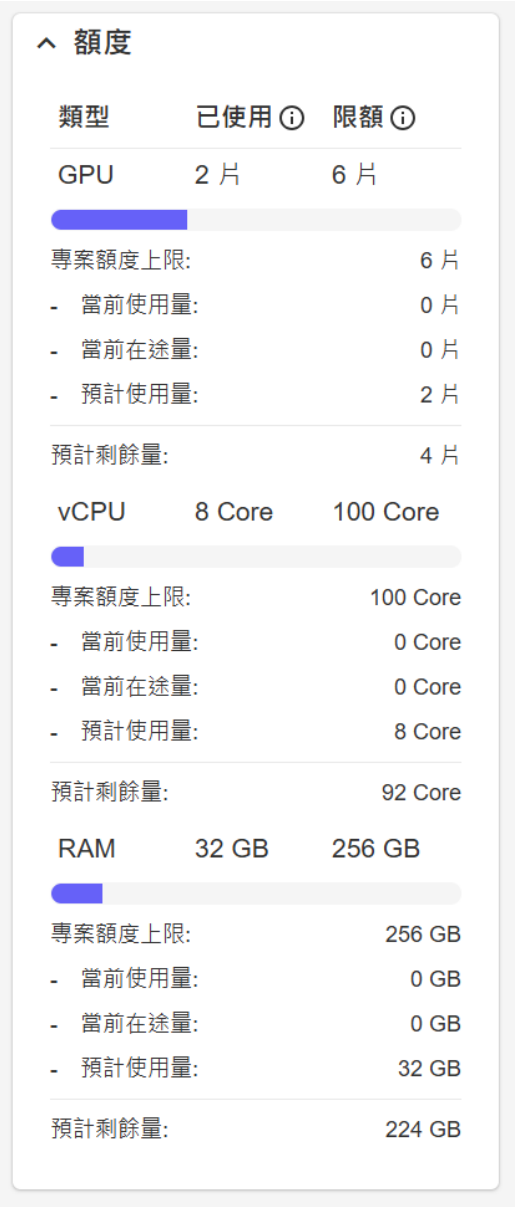
- 依照叢集類型選擇帶有訓練框架的**鏡像**，輸入要設定的服務密碼（例：SSH）。



- **啟用 GPU** 預設開啟。
- 依照訓練需求選擇**規格**。若規格選擇 Nvidia 的 GPU，請選擇 **CUDA 版本**。

  選擇規格後，最上方的**額度**會顯示目前所選規格所使用額度，可點按 › 展開 GPU、vCPU、RAM 詳細使用資訊，包含當前使用量、額度上限與額度剩餘量。

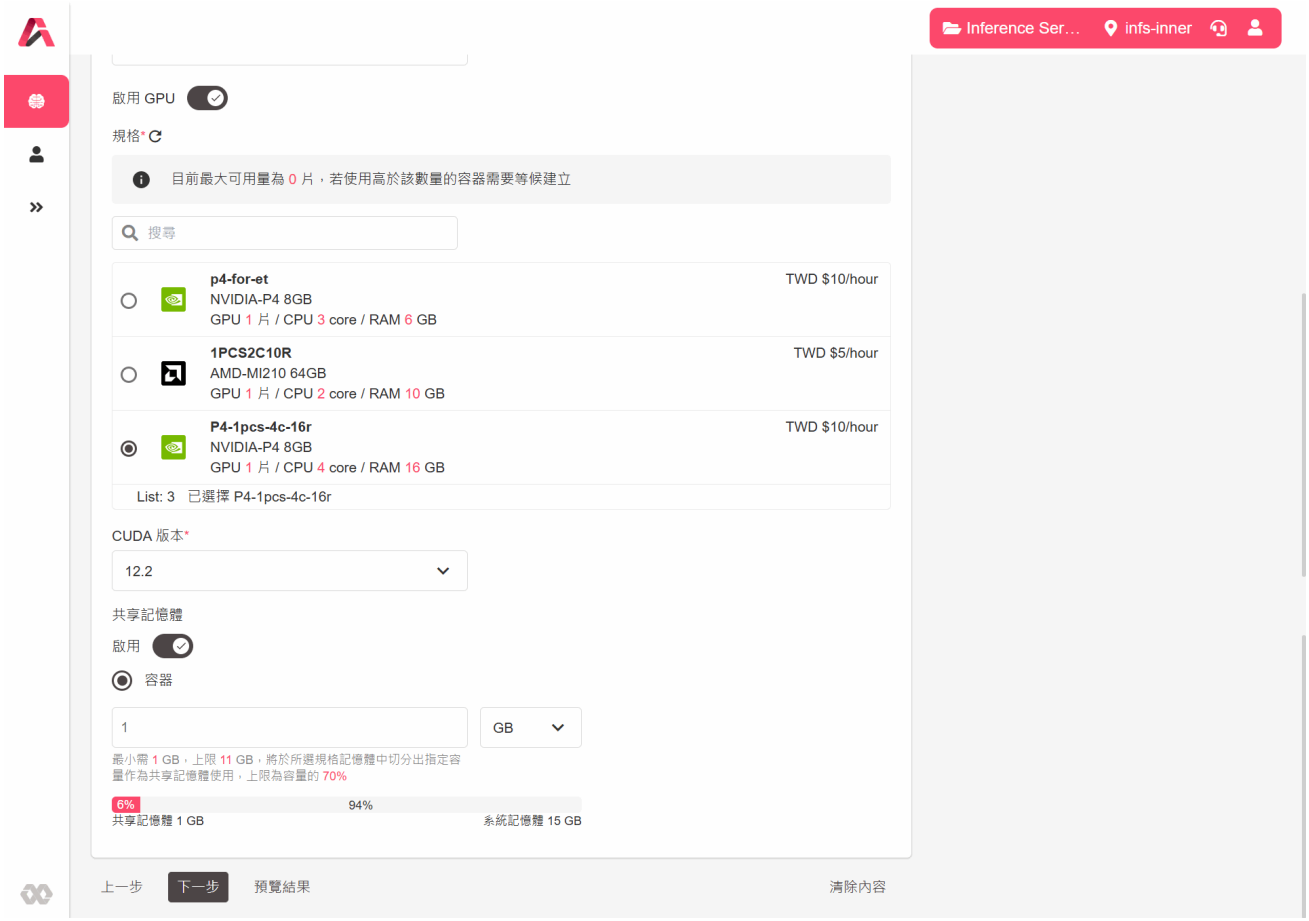




- **共享記憶體**預設開啟，並建議設定至少 1 GB 容量。

  共享記憶體用於程序間交換資料，啟用共享記憶體可加速多 GPU 訓練速度，實際容量需求依訓練資料量及訓練方式而異。

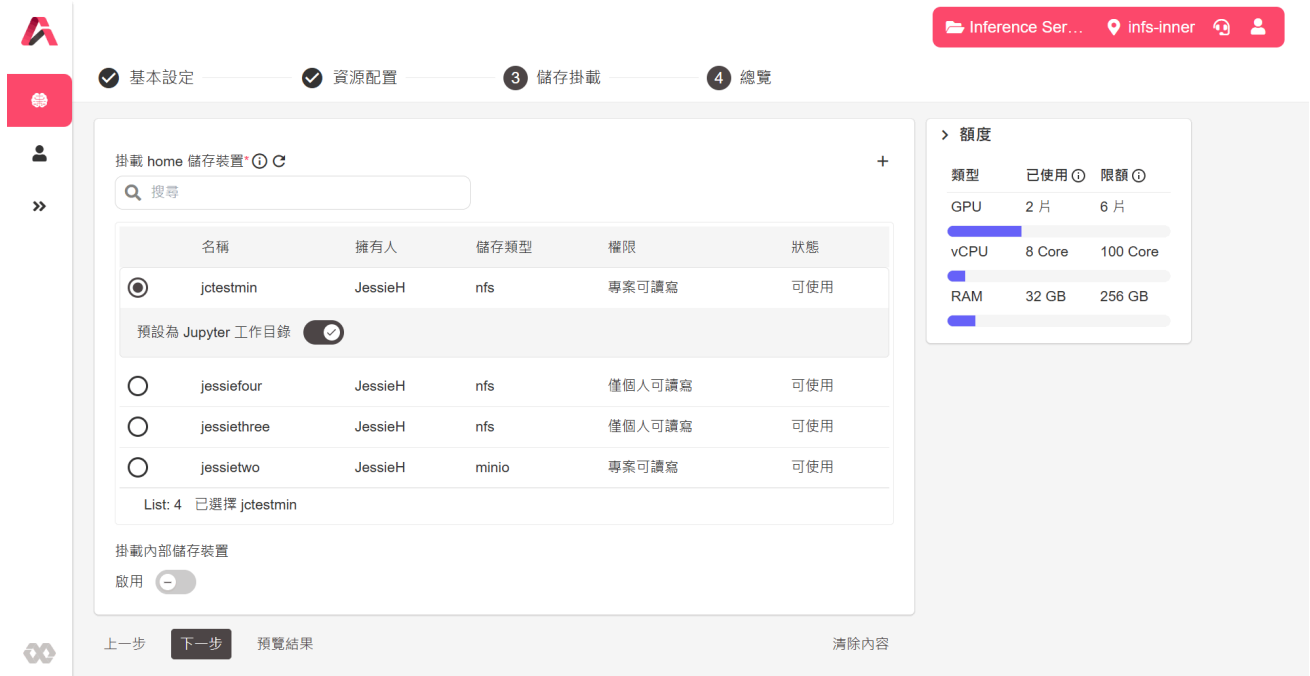
- 點擊 [下一步] 設定儲存掛載。



③ 儲存掛載

- 設定**掛載 home 儲存裝置**，選擇要掛載的裝置，若需要可啟用 [預設為 Jupyter 工作目錄]。

  * 若需新增儲存裝置，請點按右方 ＋ 或 [新增儲存裝置] 按鈕。



Home 儲存掛載說明

為了最佳化分佈式訓練流程與提升開發體驗，本平台將使用者的 Home 目錄掛載至訓練環境中。掛載 Home 儲存裝置可以讓您：

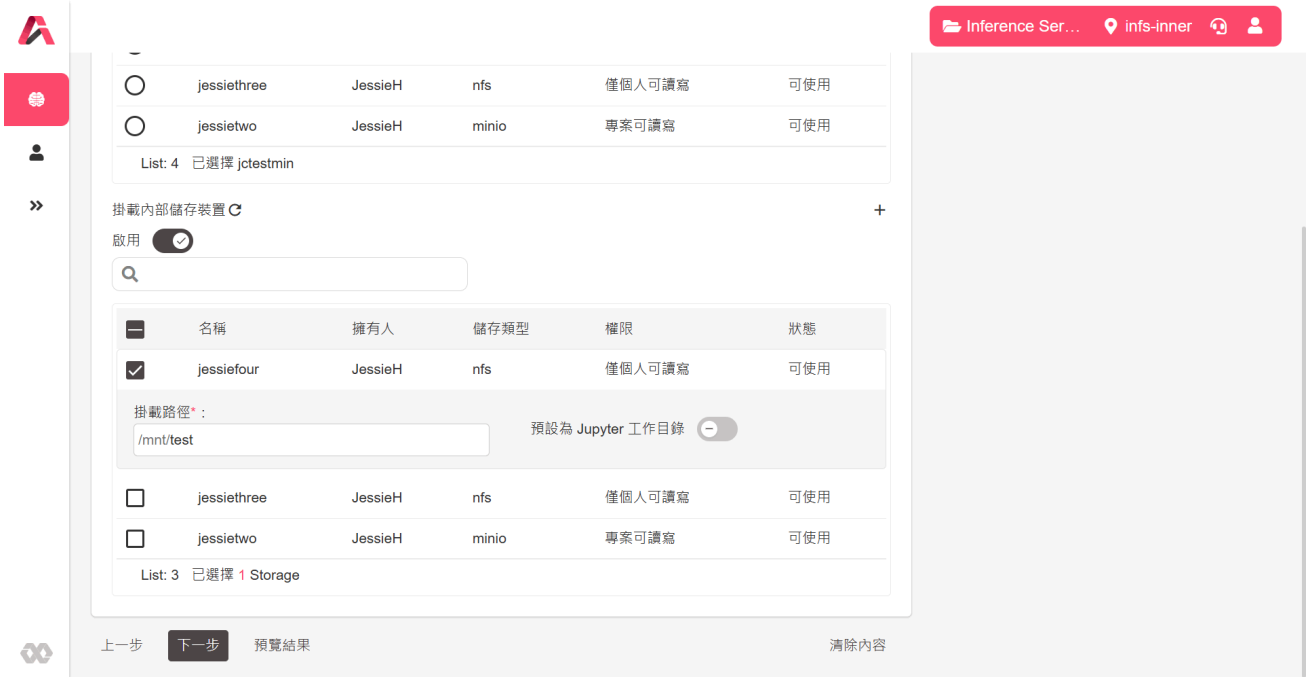
- **保留已下載的模型與快取檔案**（如 Hugging Face .cache 目錄），避免每次訓練重新下載，節省大量時間與頻寬。
- **保存個人化環境設定**，例如自訂的 Python 套件、指令稿、Alias 設定等，可隨容器啟動自動套用。
- **持久化訓練過程中產生的暫存資料**，如 DeepSpeed 的 checkpoint、optimizer state，提升中斷復原的便利性。
- **加速分佈式訓練的同步流程**，便於節點間共享初始化設定或中介結果。

平台會將每位使用者的 Home 資料夾自動掛載至訓練叢集的工作節點中，確保環境一致性與資料持續性。建議在模型訓練與推論過程中，將大型資源檔案、快取資料與個人設定統一管理於 Home 目錄內，以充分利用此功能。

* 注意：

1. 請勿在 Home 目錄儲存大量短期無需保留的暫存檔案，建議將臨時檔案儲存於暫存空間中，以維持儲存效率。
2. 訓練叢集刪除後，此目錄下的檔案並不會自動刪除。

- 若需要，啟用**掛載內部儲存裝置**並設定掛載路徑。

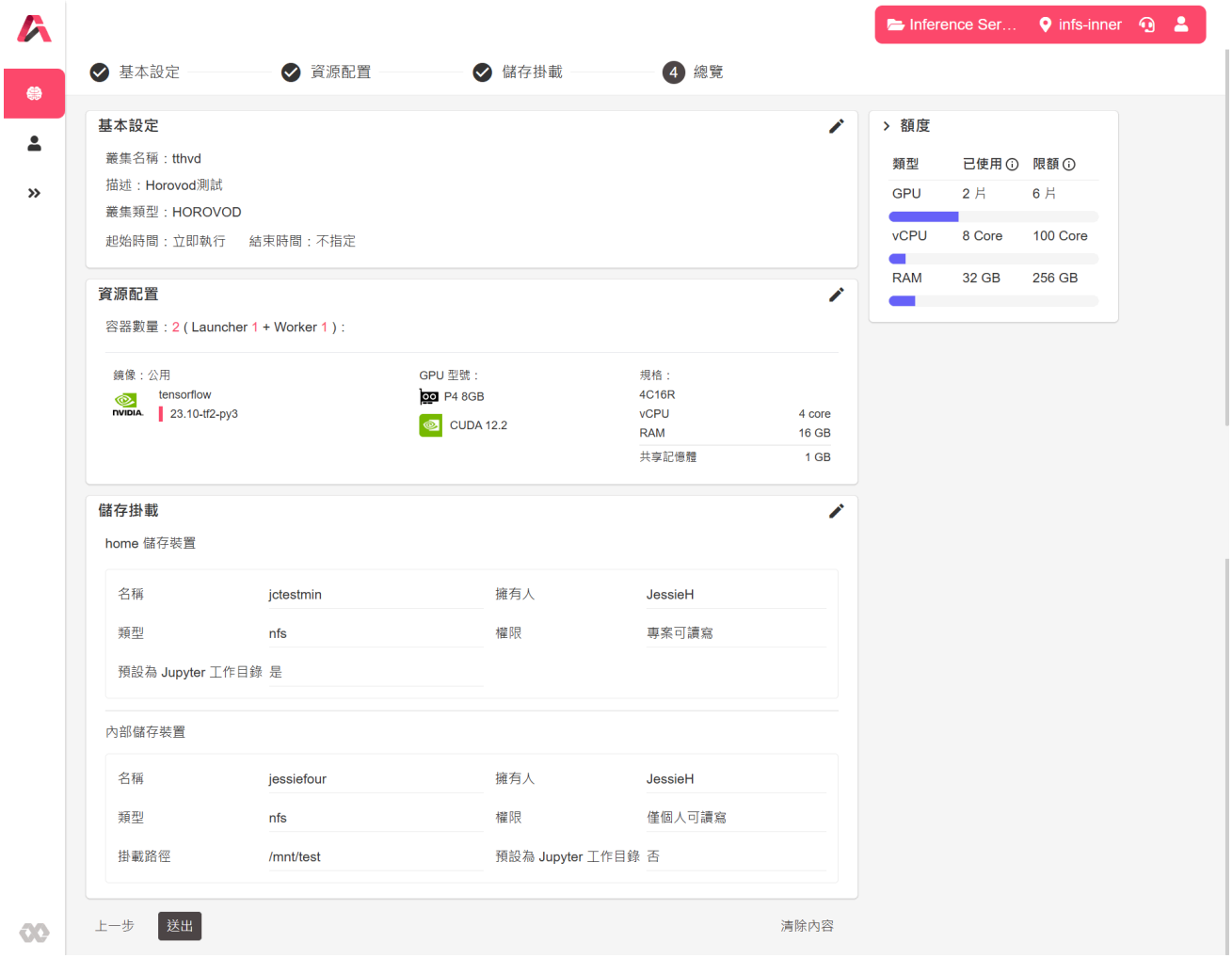


＊有關儲存裝置的相關設定，請參閱 5.7 儲存管理。

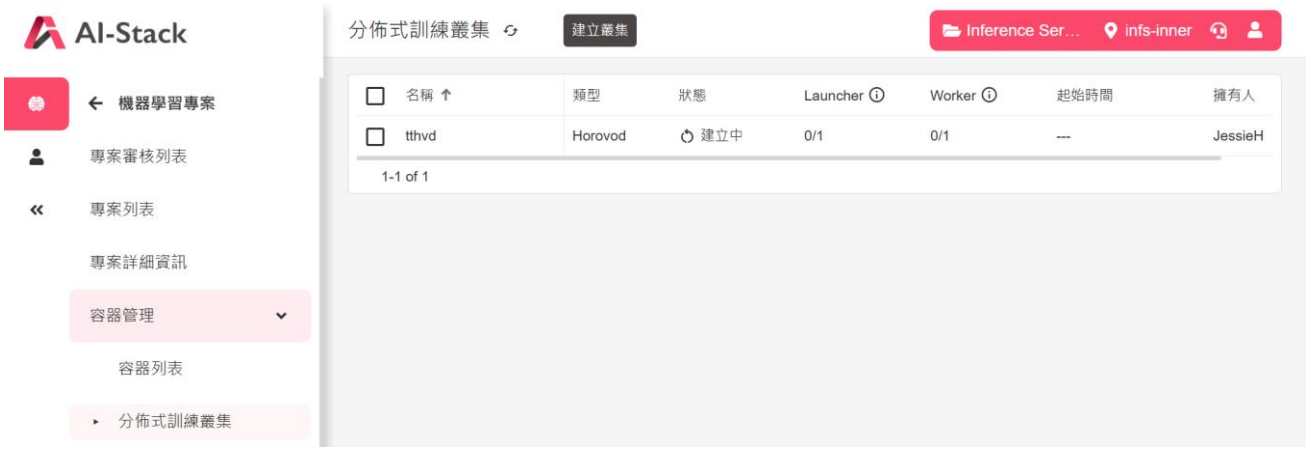
- 點擊 [下一步] 進入總覽頁。

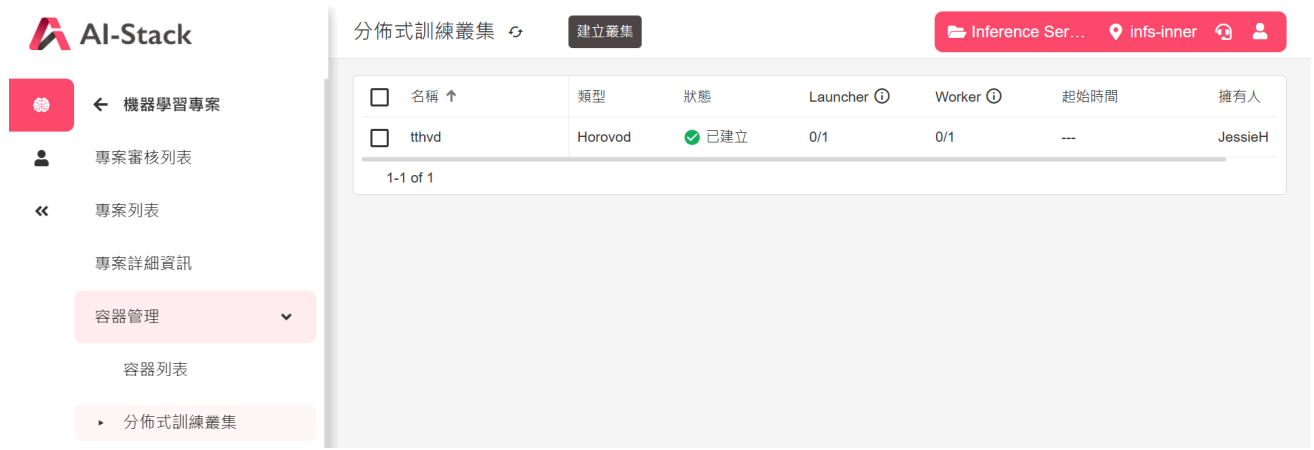
④ 總覽

- 確認設定無誤後，點擊 [送出]。若需修改內容，點擊右上角 回到該步驟編輯。



- 送出後會回到【分佈式訓練叢集】列表，待叢集狀態由「建立中」更新為「已建立」，分佈式訓練叢集即建立完成。

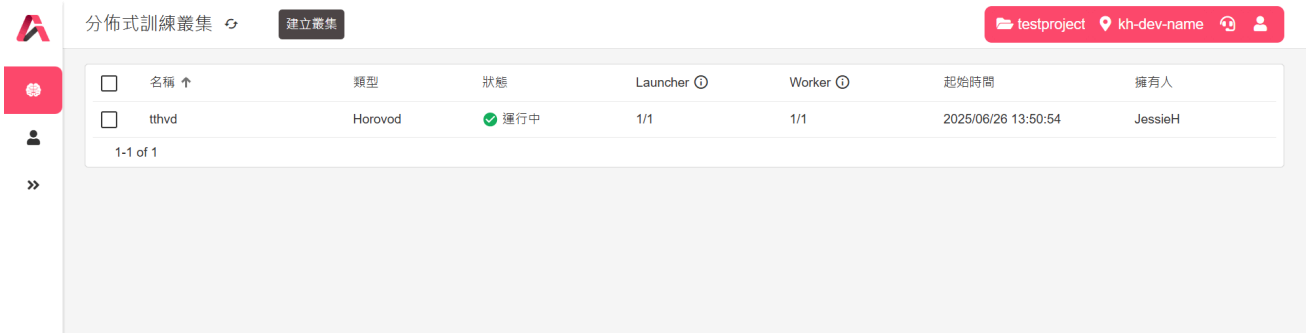




6.2 分佈式訓練叢集列表

6.2.1 叢集列表

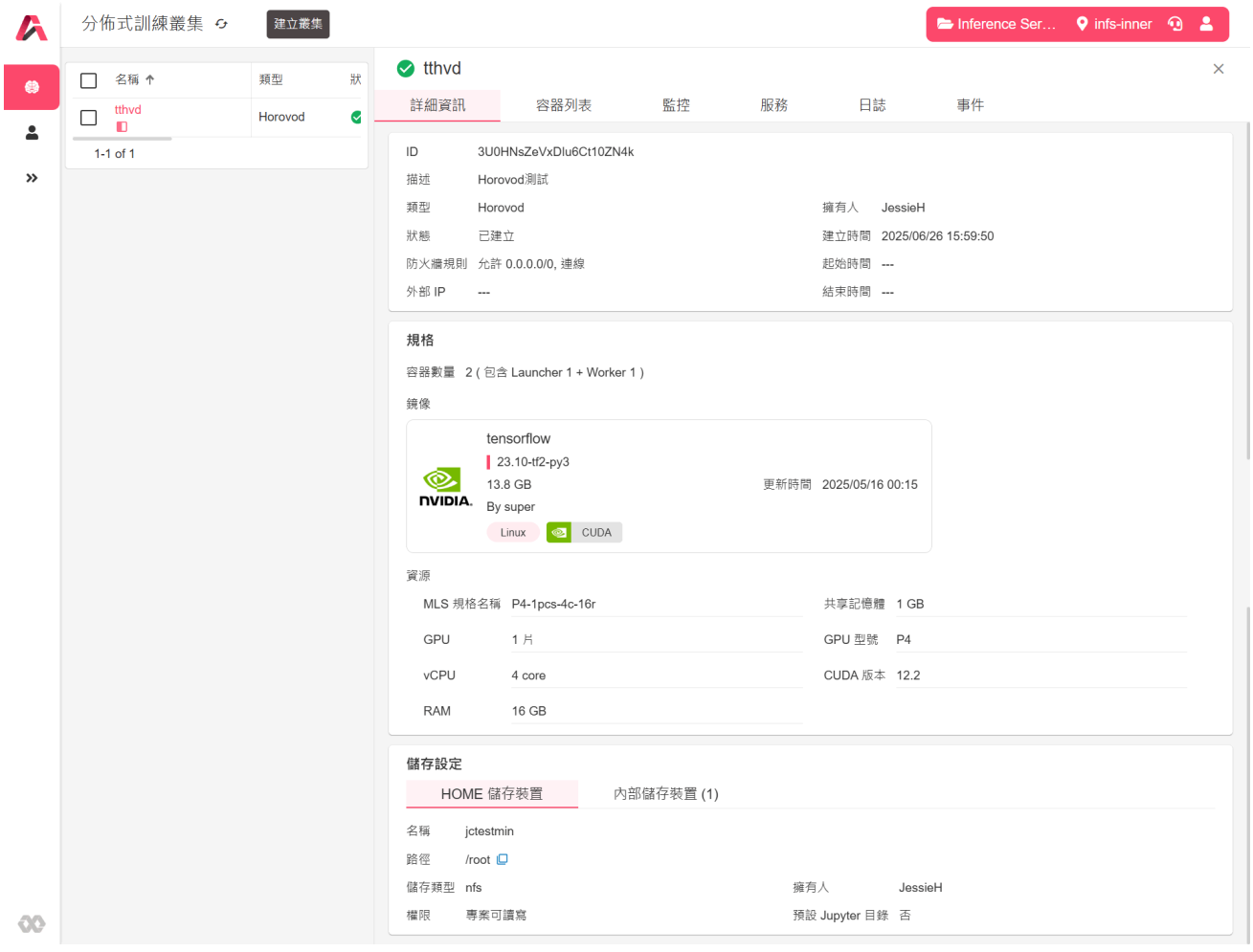
【分佈式訓練叢集】頁列出已建立的容器叢集，包含叢集類型、運行狀態、Launcher 容器及 Worker 容器數量（已建立 / 總數量），以及叢集運行起始時間。



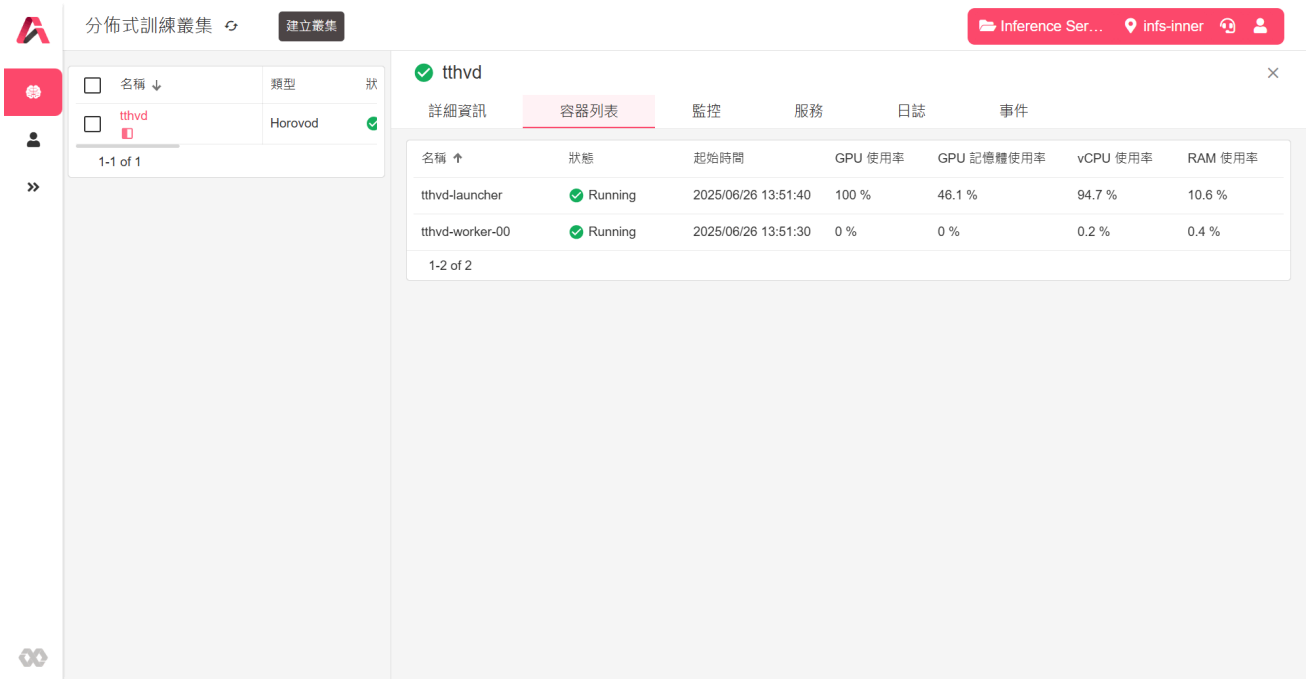
6.2.2 叢集詳細資訊

在【分佈式訓練叢集】頁選擇欲查看的容器叢集，即可查看下列詳細資訊：

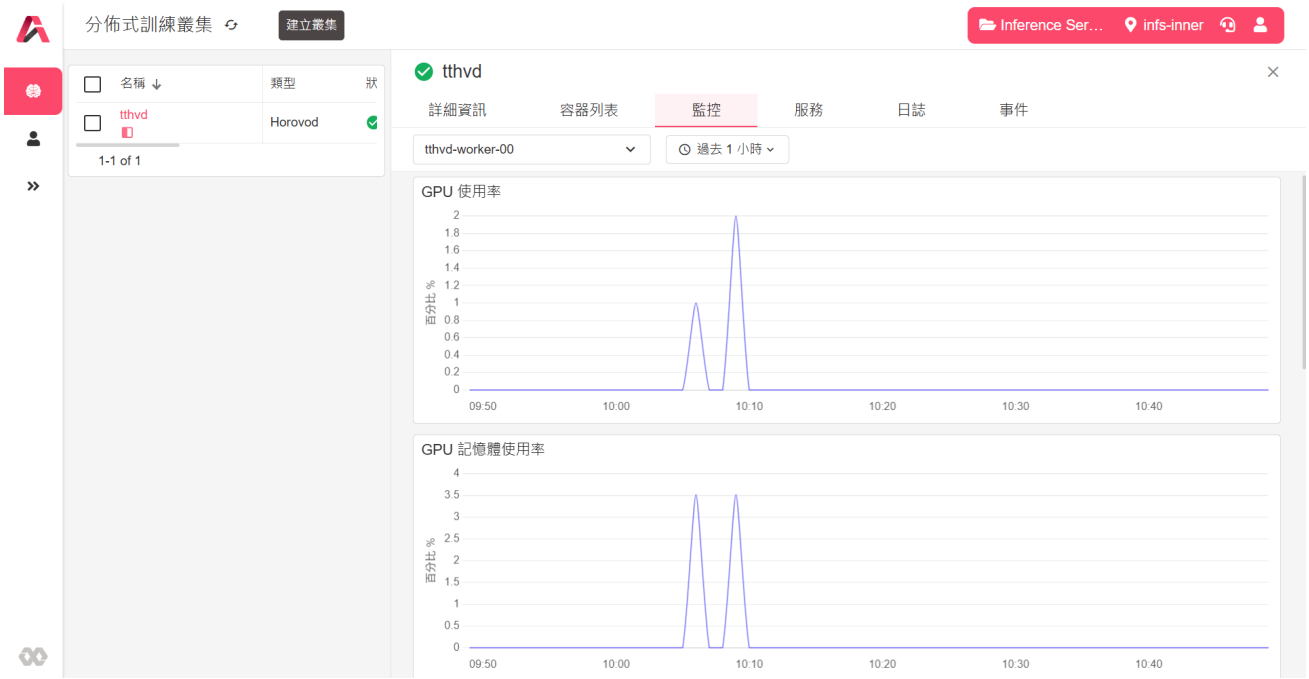
- [詳細資訊]：可查看容器叢集基本資料、容器數量、所選鏡像與 MLS 規格及儲存裝置設定。

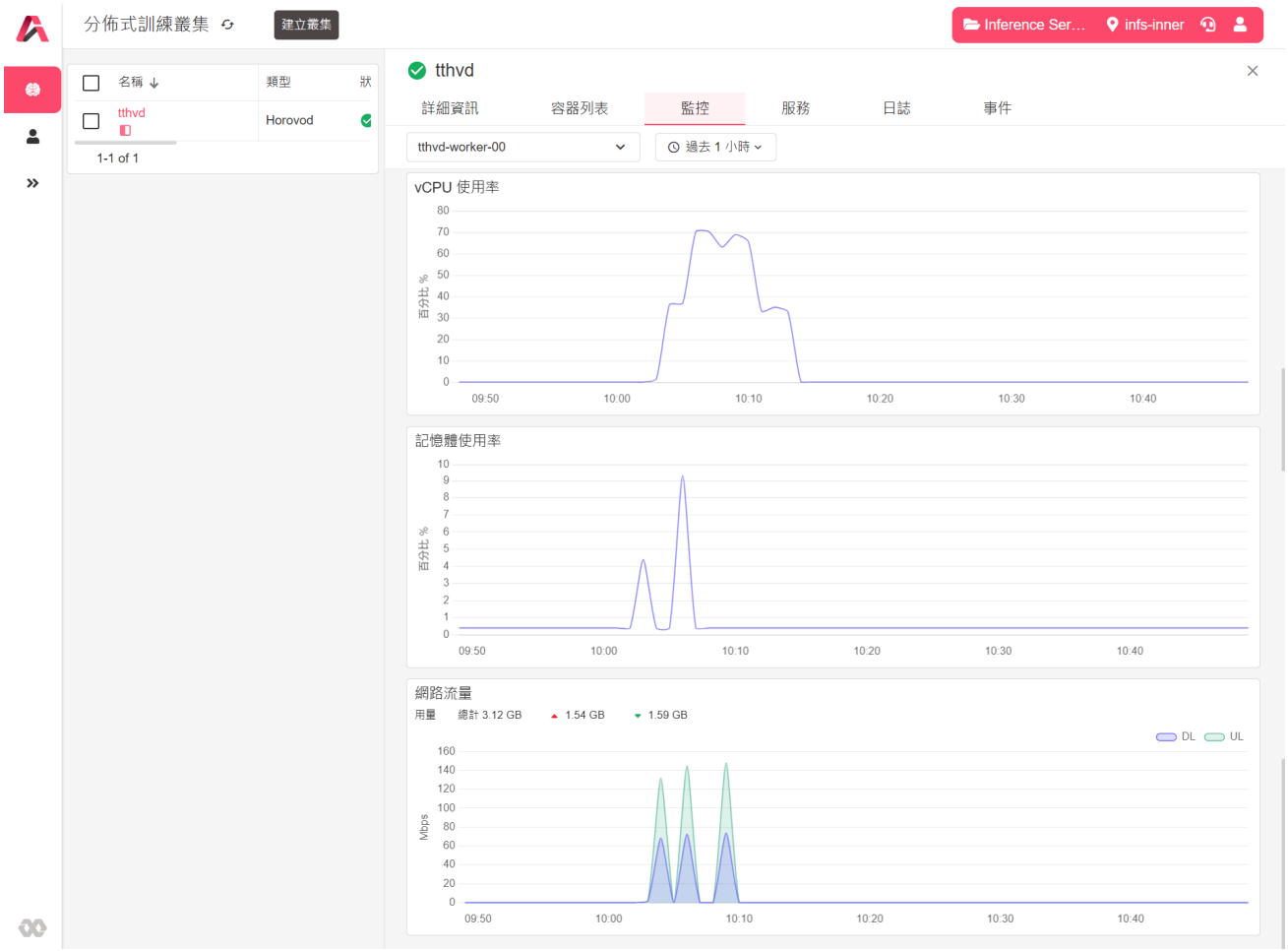


- [容器列表]：可查看叢集內所有的容器及目前運行資訊，包含運行狀態、起始時間、GPU 使用率、GPU 記憶體使用率、vCPU 使用率、RAM 使用率。

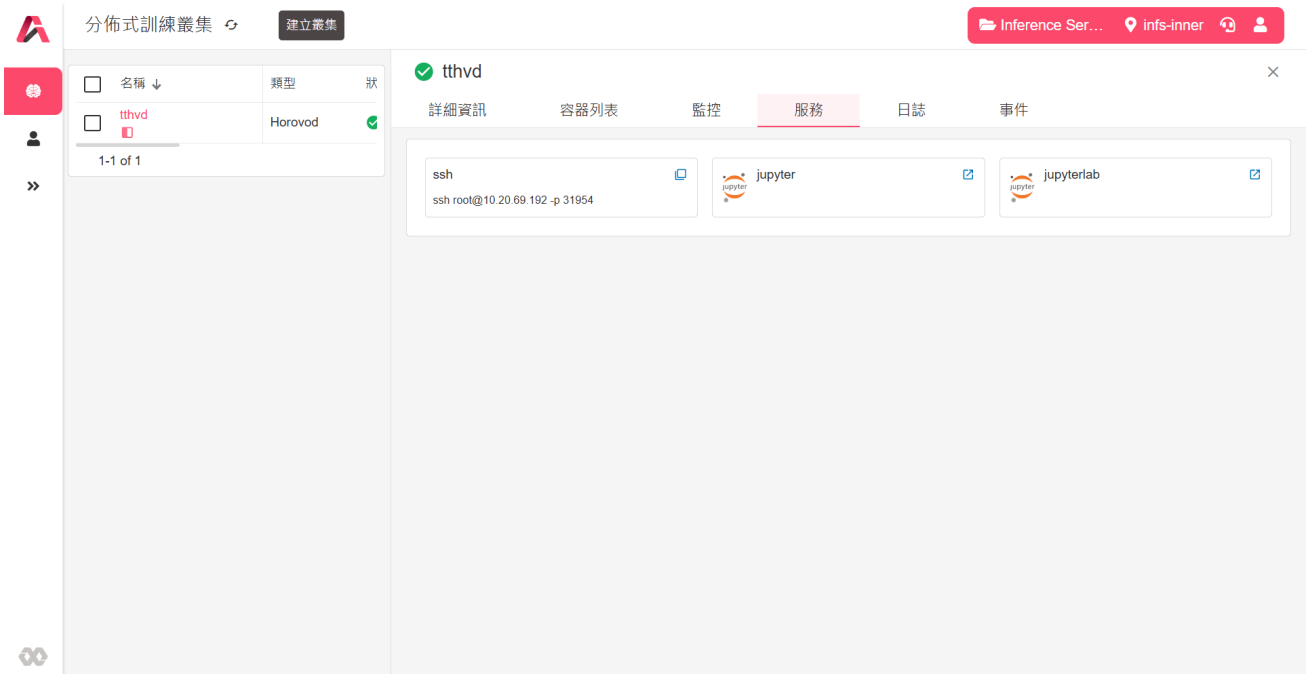


- [監控]：可查看單一容器在指定時間內的詳細監控資料，包含 GPU 使用率、GPU 記憶體使用率、vCPU 使用率、RAM 使用率及網路流量。可透過下拉選單選擇要查看的容器及指定的時間，最多查詢區間為 30 天。

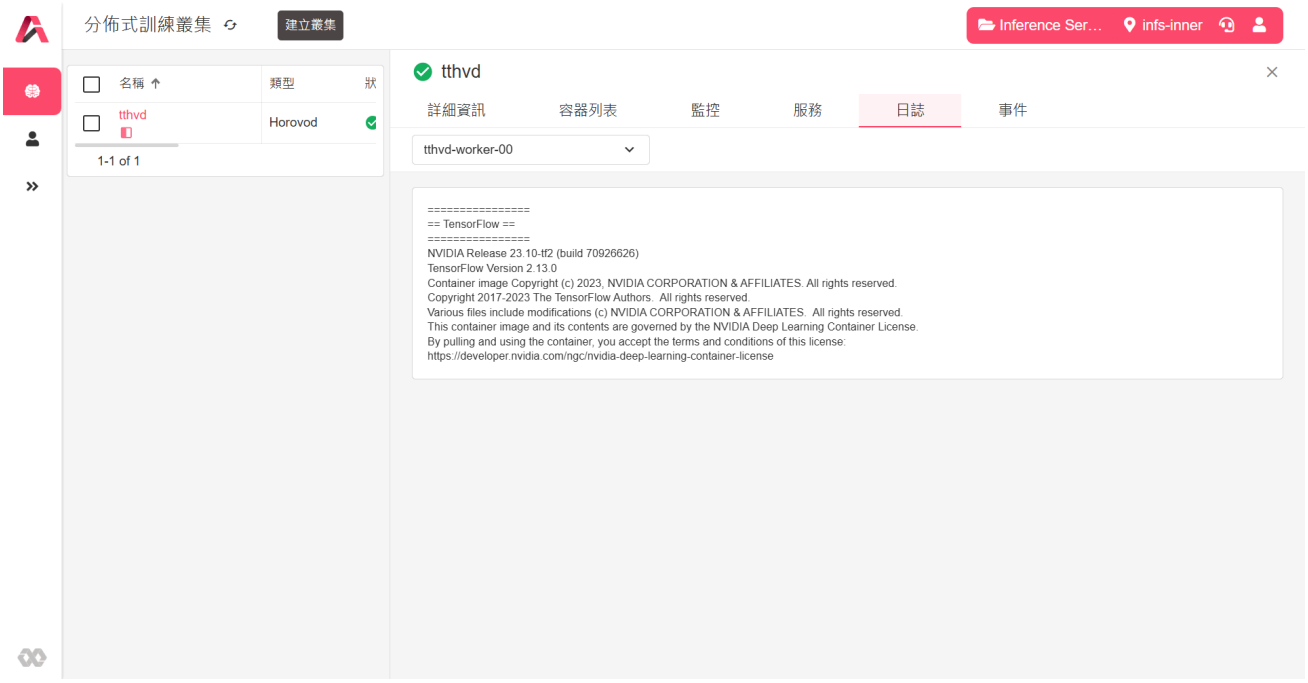




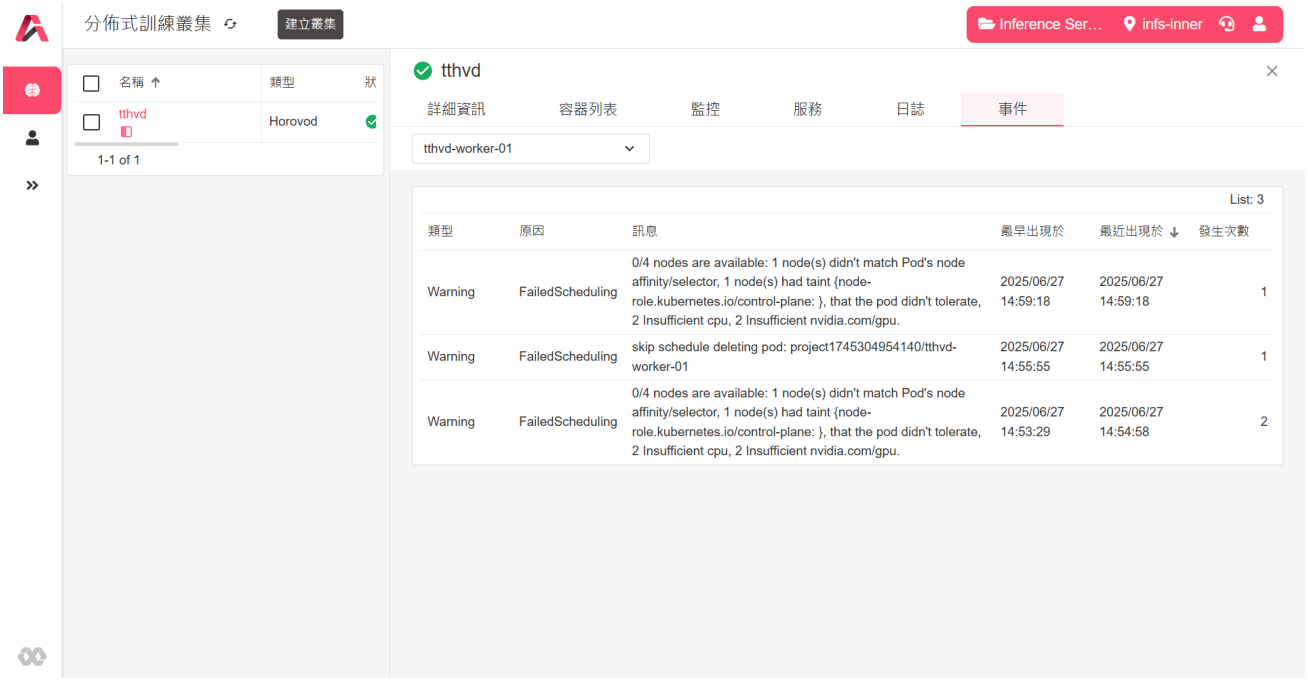
- [服務]：可查看叢集內提供的所有服務，服務及啟用方式由平台管理者透過 MLS 樣板（鏡像）定義，支援 SSH、Jupyter、JupyterLab、TensorBoard、WebTerminal、Code Server 與使用者自行定義。



- [日誌]：可查看單一容器在運行過程中的所有日誌輸出紀錄，包括系統訊息、應用程式運行紀錄及錯誤報告等詳細內容，有助於用戶掌握容器的實際運行情況、追蹤執行過程中的關鍵步驟，並進行問題診斷與除錯。可透過下拉選單選擇要查看的容器。

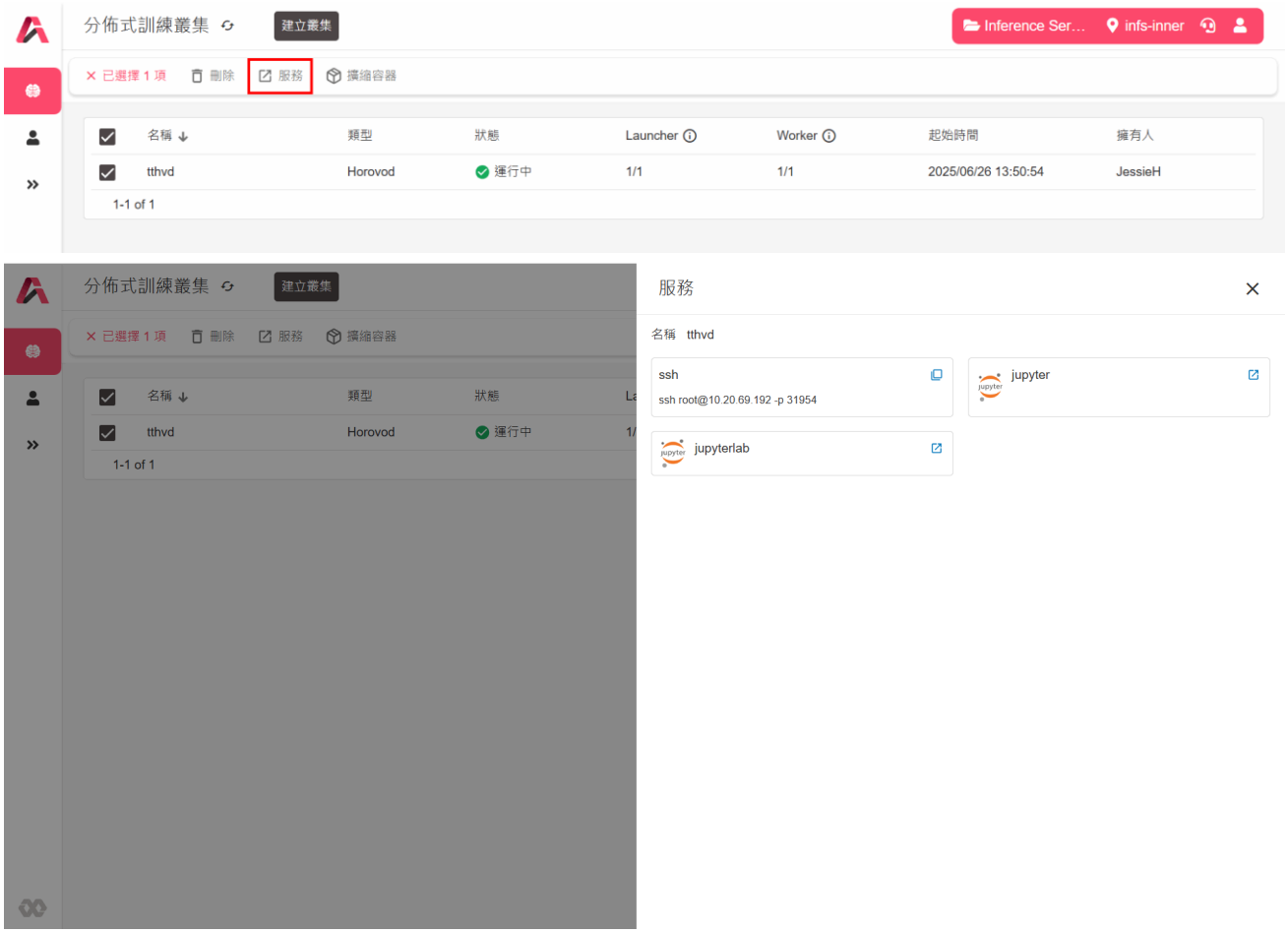


- [事件]：可查看單一容器在運行過程中發生的重要狀態變化或觸發事件，事件紀錄幫助用戶瞭解容器狀態變化，進行異常排查與故障追蹤。可透過下拉選單選擇要查看的容器。



6.2.3 叢集服務

在【分佈式訓練叢集】頁，選擇欲查看的叢集後，點擊 [服務]，即可快速開啟本容器叢集的服務頁，服務及啟用方式由平台管理者透過 MLS 樣板（鏡像）定義。



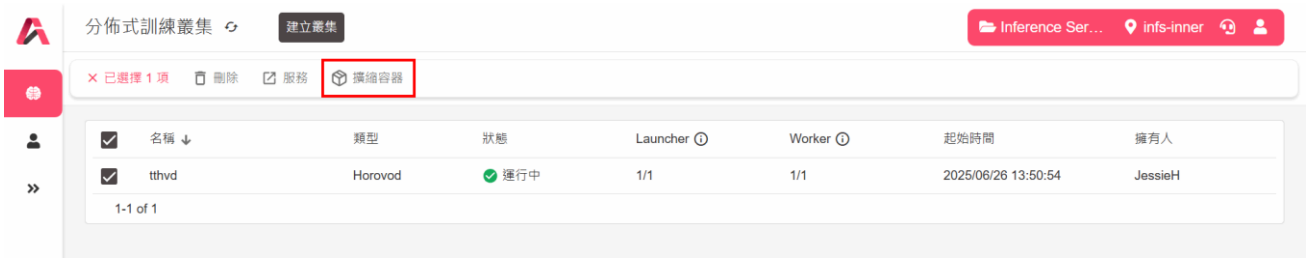
6.2.4 擴縮容器數量

6.2.4.1 擴展容器數量

在叢集使用過程，如果需要更多運算資源，可隨時擴充叢集容器數量。步驟如下：

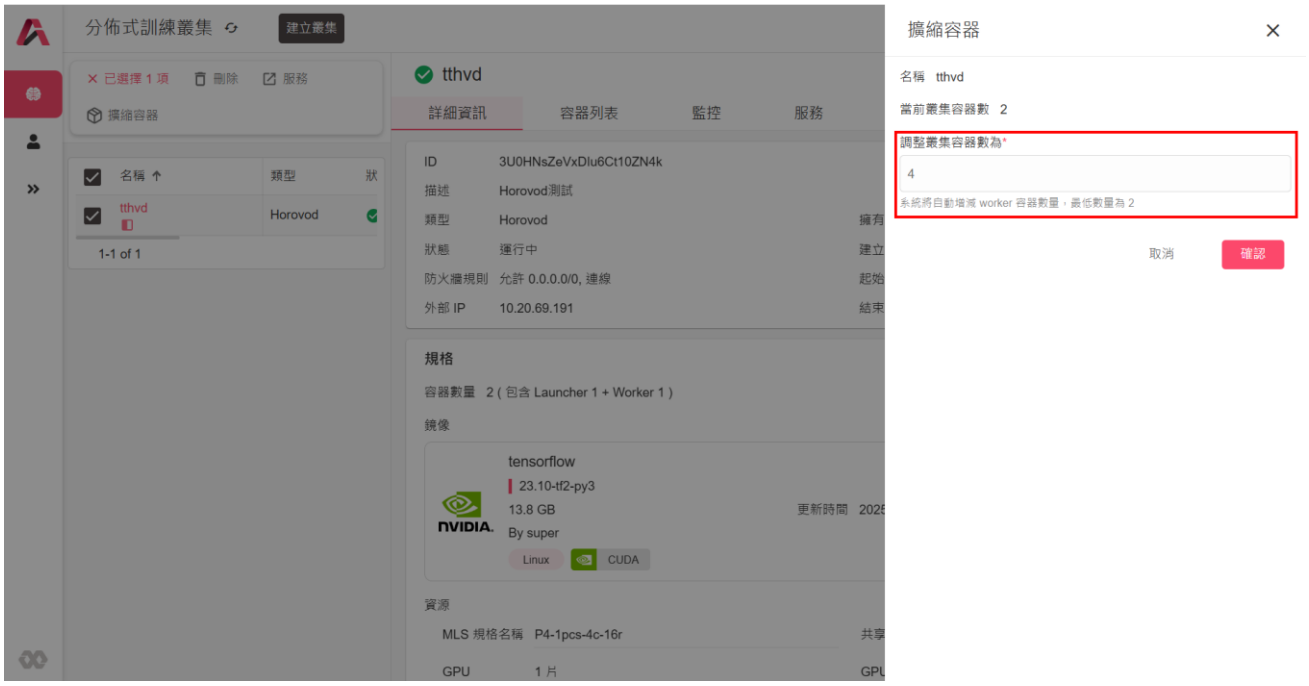
- 在【分佈式訓練叢集】列表中，選擇需要擴展的容器叢集，點擊上方 [擴縮容器] 按鈕。

  * 注意：執行擴容時，建議停止運行腳本，再進行擴容設定。

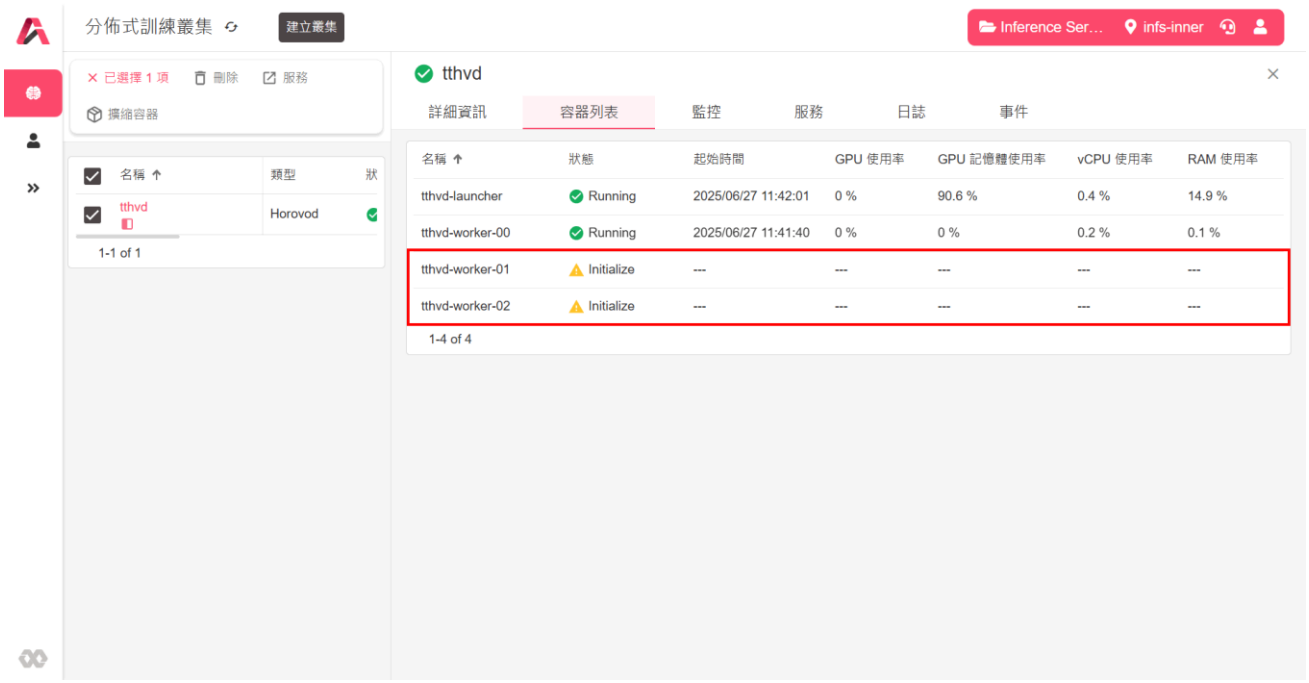


- 在 [調整容器數量為] 欄位中輸入新的容器數量，此數值是總數量，包含目前已經建立的 worker 容器及 1 個 launcher 容器，最低數量為 2。系統將自動增加 worker 數量容器並修改 hostfile。

  Hostfile 自動修改的容器數量會與實際運行的容器數量一致。若因資源不足導致容器仍在等待建置中，則 hostfile 不會跟著修改，避免運行腳本時產生錯誤。



- 確認數量後，點擊 [確認]。
- 回到 [容器列表]，可以看到剛剛擴增的容器，待容器狀態更新為「運行中」，代表叢集擴展、啟動完成。

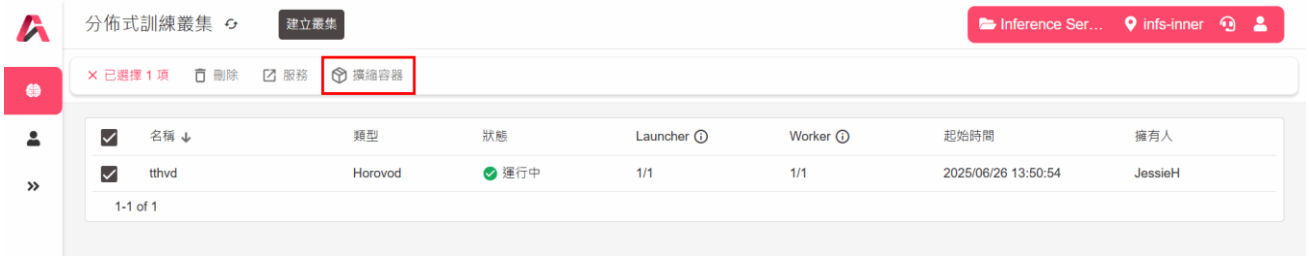


6.2.4.2 縮減容器數量

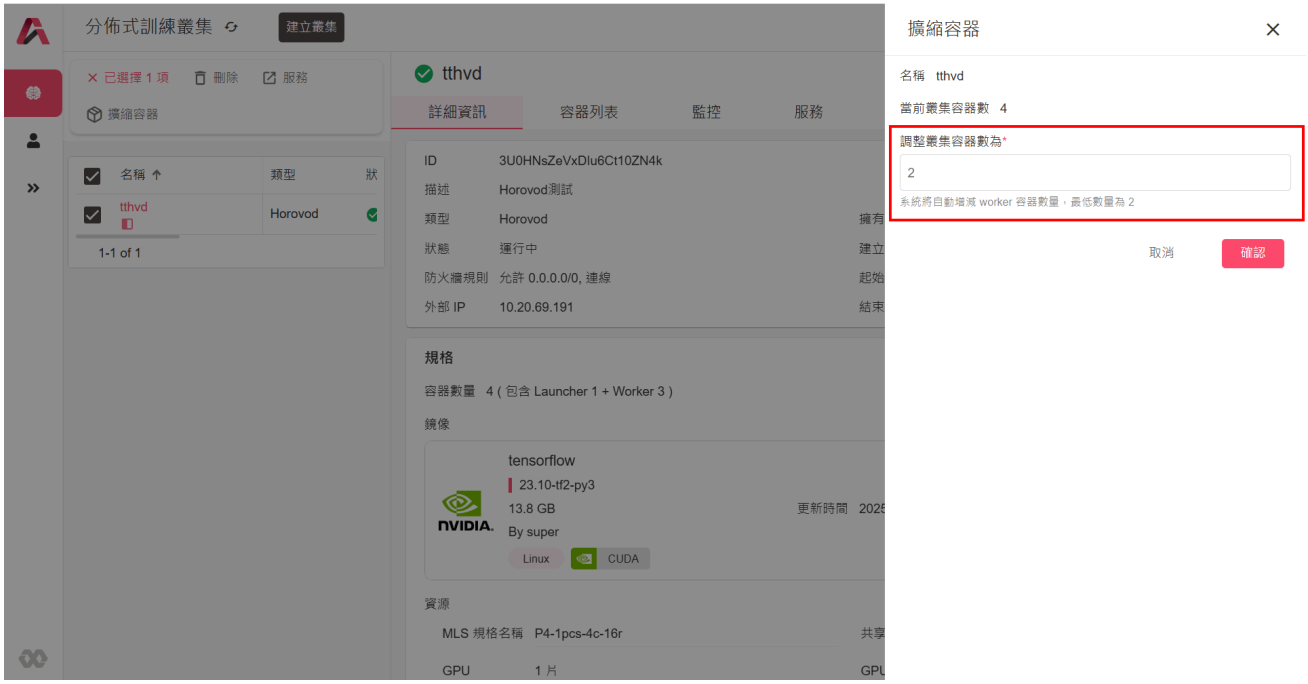
當訓練叢集不需要這麼多資源時，可根據實際需求刪除部份容器，將資源釋放出來，步驟如下：

- 在【分佈式訓練叢集】列表中，選擇需要縮減的容器叢集，點擊上方 [擴縮容器] 按鈕。

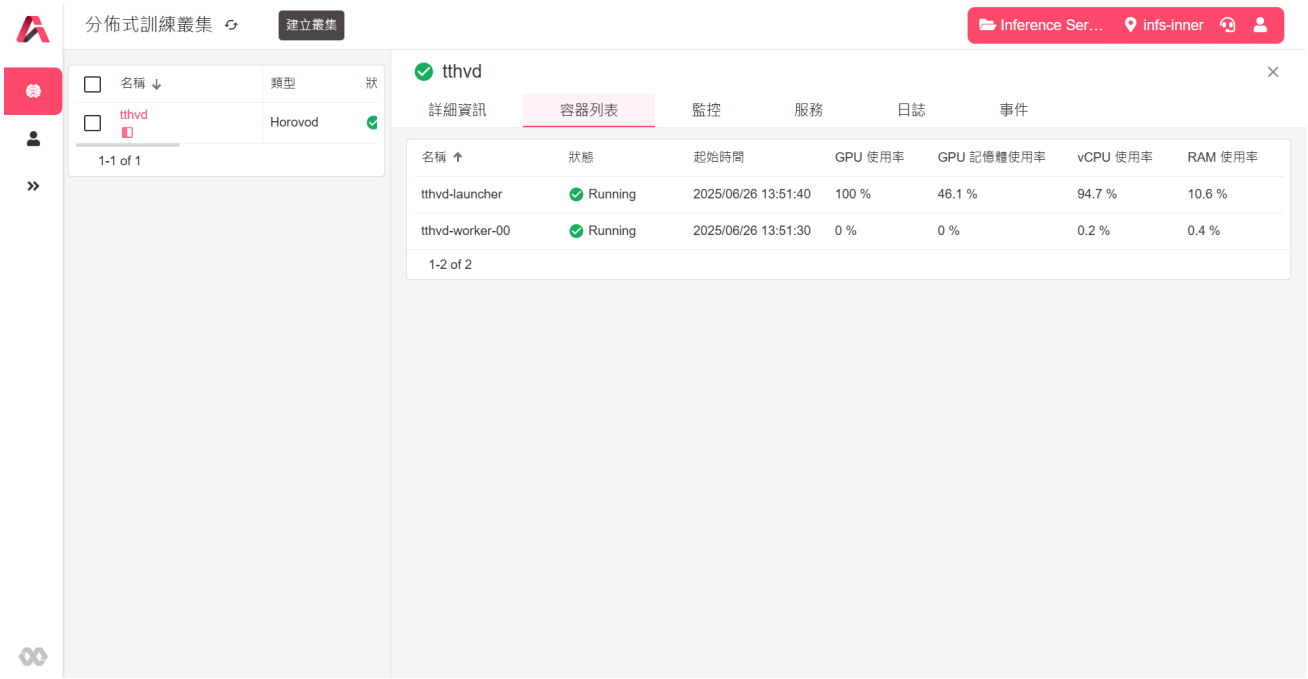
  * 注意：執行縮容時，建議停止運行腳本，再進行縮容設定。



- 在 [調整容器數量為] 欄位中輸入新的容器數量，此數值是總數量，包含目前已經建立的 worker 容器及 1 個 launcher 容器，最低數量為 2。系統將自動縮減 worker 數量容器並修改 hostfile。



- 確認數量後，點擊 [確認]。
- 回到 [容器列表]，等待縮減的容器從列表中刪除，代表叢集縮減完成。



6.2.5 刪除叢集

在【分佈式訓練叢集】頁，選擇欲刪除的容器叢集（可多選），點擊 [刪除] 後將出現確認視窗，確認所選為想要刪除的容器後，再次點擊 [刪除]。