


透過數據及AI助力 行業轉型與加速創新

- 技術副總經理- 台灣戴爾科技集團
- 梁匯華/ Eric Leung

 @ar_wa2000



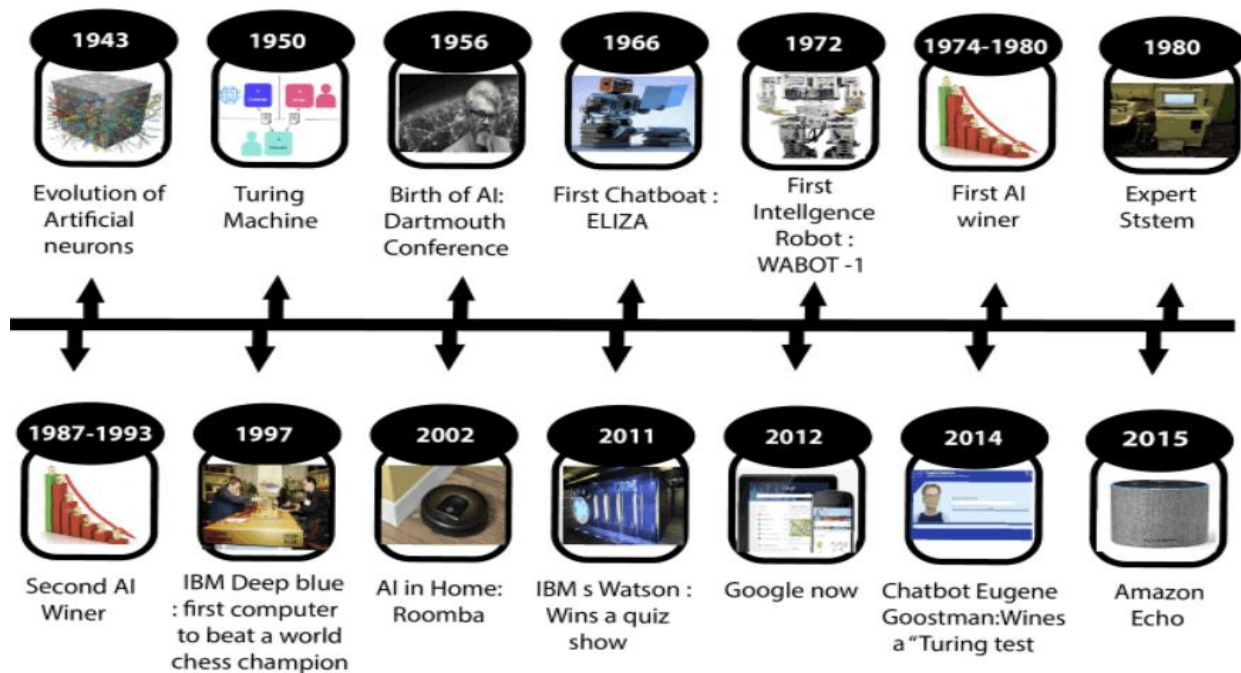


第一次的AI 對奕

Bertie the Brain pitted humans against an artificial intelligence in a game of tic-tac-toe

人工智慧史

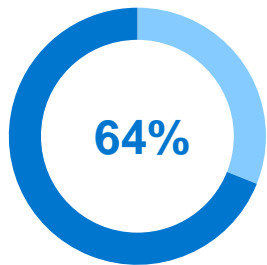
人工智慧歷史上的一些里程碑



- 人工智慧不是一個新詞，也不是研究人員的新技術。這項技術比你想像的要古老得多；甚至還有古希臘和埃及神話中機械人的神話。

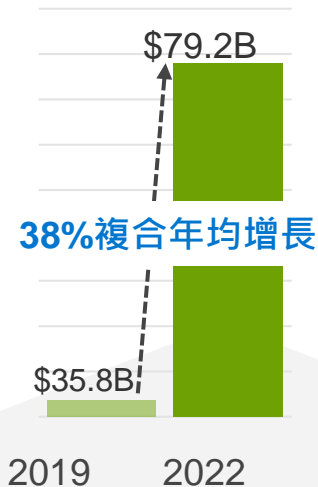
我們正處於AI發展的臨界點

誰會參與其中？



在未來3年中，CIO將投資於機器學習領域

全球在人工智慧系統上的支出



交易規模

整體方案

- ~\$265K Ready Solutions for AI avg
- ~\$200K for a proof-of-concept configuration

AI 優化儲存設備

- ~\$250K starting point
- ~\$600K average

伺服器

- ~\$5K workstation starting point
- ~\$65K 4-GPU server
- \$+++ growth

¹ ServiceNow: [The Global CIO Point of View](#), 2019

² IDC. Press Release: "[Worldwide Spending on Artificial Intelligence Systems Will Grow to Nearly \\$35.8 Billion in 2019.](#)"

人工智慧的未來就是現在

一場完美的風暴，推動著客戶的需求

10101
11010
01011

更多
數據

有越來越多的數據可用於為人工智慧-每秒生成更多數據。其中大部分是數據為非結構化數據。



大大增強的
計算力

多線程 GPU 現在提供電源演演算法，即時處理，便於快速識別趨勢和模式。



AI
創新

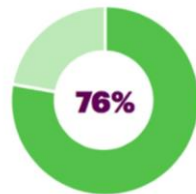
現在，我們可以訓練機器使用數據來感知、學習、推理、預測和創新之應用。

智慧保險顧問催生AI 應用

智慧保險顧問指的是奠基於人工智慧及大數據分析技術，依據顧客基本資訊提供個人化風險分析、投保建議的創新型保險服務

- 相對較公正客觀
- 全面掌握顧客需求
- 不受時空限制

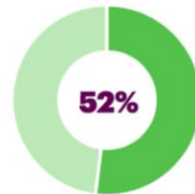
數位化也促使新生代客群對保險公司的依賴程度下降。全球市場上，不只有更多的顧客傾向使用自動化工具獲取投保建議，新生代客群當中，更有將近9成在購買保險時選擇聽取朋友或家人的意見，而不是保險公司的建議



期待保險公司能夠針對住在家裡的老人提供健康生活保障



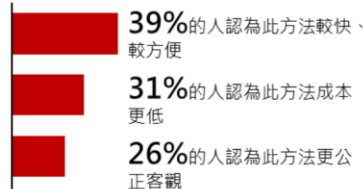
期待保險公司能夠根據所在位置提供客製化的保險服務



期待保險公司能夠提供行動裝置以遠端遙控智慧家居



購買保險時願意接受電腦產生的建議



全球市場上顧客對於自動化投保建議採取開放態度(資料來源：Accenture 2017市場調查)

案例1: 產品外觀檢測

Dell 筆電 Latitude A cover 組裝線

[AI 機器視覺識別]

TRIVISION AI 模型, 篩查十大類常見缺陷, 並告知缺陷類別和位置以及可能成因, 並繼續學習新缺陷, 其中:

- 灰度演算法, 篩查嚴重不良;
- 比良演算法, 篩查未經學習的缺陷, 比對未經學習過的缺陷, 不良防堵以防漏失。

Data scale: **每10秒**完成一個 A cover組裝, 產生**250MB**數據, **1天產生2TB**數據。資料需保留**2年**, 約**1PB**數據儲存在 **Isilon**上

[具體成效]

- ✓ 該環節的良品率 yield rate 提升 **10%+**
- ✓ 漏殺率 missing rate (把不良品誤判為良品) **從 0.5% 到 0.3%**;
- ✓ 過殺率 overkill rate (把良品判為不良) **從 5% 到 3%**;
- ✓ 節省組裝線人力配置: 一條流水線設 **5 個人工**檢測站, 機構件廠還需要外派人員到ODM做入料篩查和隨線篩查。



案例2: 產品組裝件檢測 - Dell 筆電組裝線

Before
手工檢測



在組裝前, 62個檢測點, 手工的一個個檢測。

After
CCD+AI 檢測



配上光學設備+AI演算法,
自動檢測

[結果]

a. inspection item.

Checking items	數量
Screws	36
Cable	11
Connectors	12
Ruber	3
Summary	62

b. in line inspection

c. CT < 30sec.

d. Coverage: 95%

e. overkill: 2%

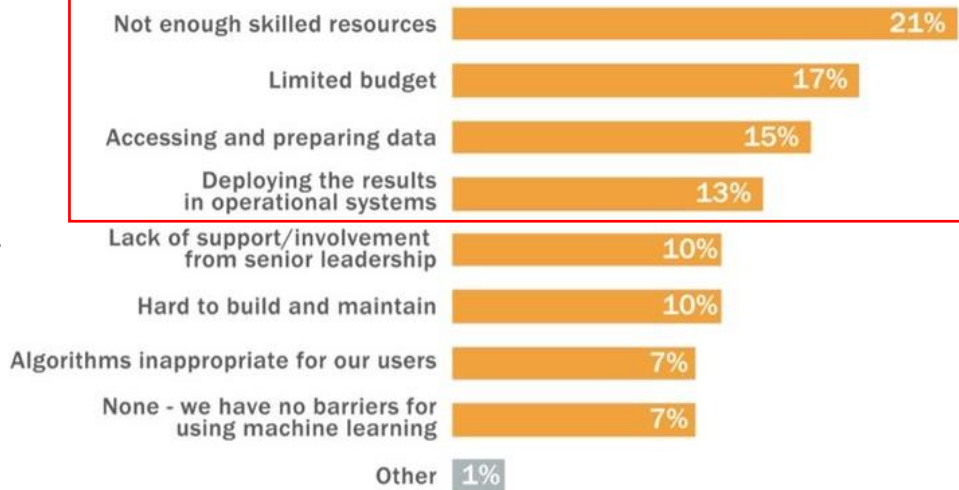
f. missing: 0.5%(by unit)

客戶對人工智慧寄予厚望



Source: 451 Research's AI & Machine Learning Use Cases 2020

然而，它面臨著重大挑戰



Source: 451 Research's Voice of the Enterprise: AI & Machine Learning, Adoption and Use Cases 2018

熊貓 vs 魚子醬

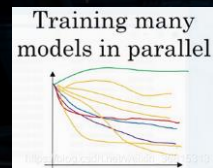
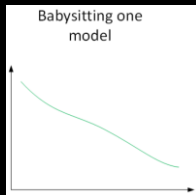


熊貓

- 在DL中，它是具有1：1比率的工作站/伺服器的個別開發人員。
- 數據科學家的主導
- 運營成本昂貴，其中包括機會成本和實際成本
- 數據大小往往被單一實體限制

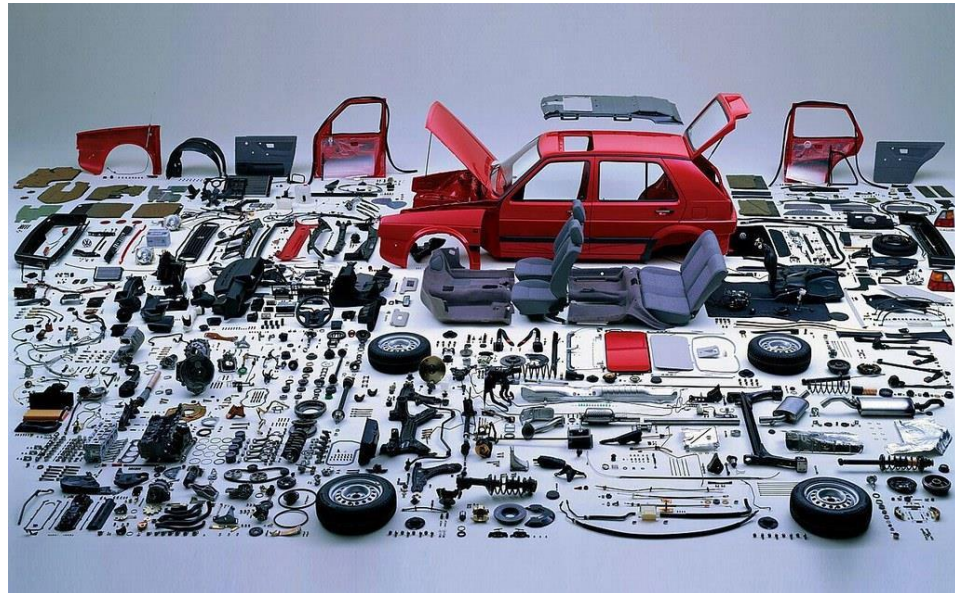
魚子醬

- 魚有上百萬個魚卵
- 如在深度學習中，構建“即服務”的基礎架構團隊
- 沒有個人擁有任何特定的“GPU”
- 由 IT 支援
- 同時顧及安全與共用資源，運營上更受歡迎
- 數據大小是彙集的，而不是孤立的；能處理的數據大小決定在整個集群有多大



企業 AI 的簡單按鈕

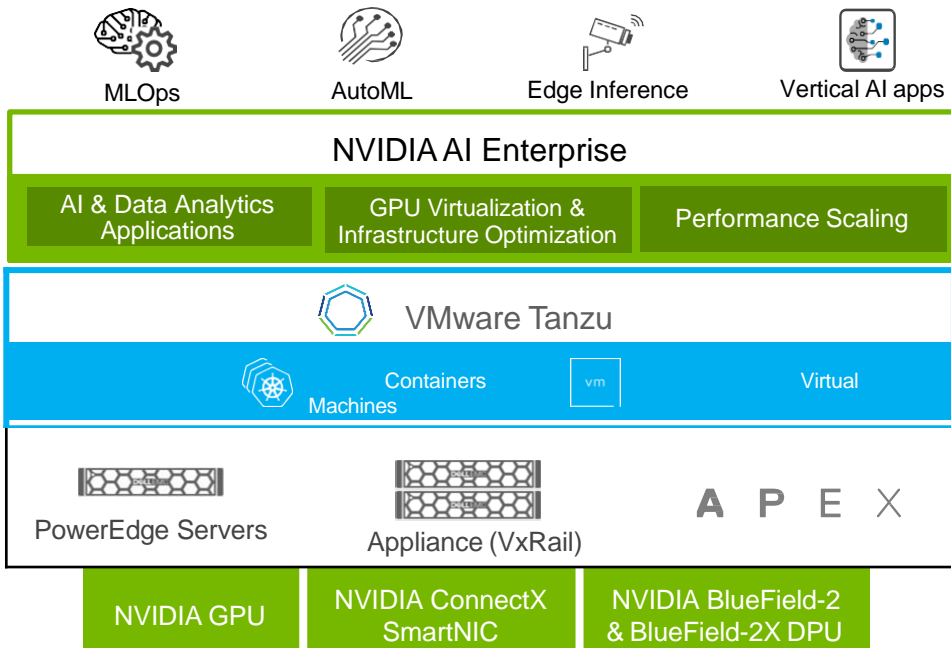
今天的DIY方法



主流人工智慧



AI Enterprise - 簡化 AI 基礎架構的部署和管理



- Full Stack Enterprise Software Product from NVIDIA
- Supports vSphere 7.0u3
- Runs on NVIDIA Certified servers from Dell Technologies
- Benefits of virtualization with bare-metal performance



IT ADMIN - BENEFITS

- Use the same tools they use to manage other enterprise apps
- Provision "right size" VMs for data scientists
- Stability of environment through release management and full-stack support from NVIDIA



AI PRACTITIONER/DATA SCIENTIST - BENEFITS

- Industry-leading AI software tools and frameworks
- Focus on data science, rather than IT management
- Guaranteed interoperability across the software stack
- Access to pre-trained models for faster model deployment time

Dell + AI-Stack加速客戶讓AI落地

AI-Stack

後端 訓練運算 (TRAINING)

運算所需要的巨量資料, 大量的圖片, 影片, 文字或語音

資料處理

訓練運算

模型評估

經人工智慧訓練運算產出的類神經網路軟體, 將安裝佈署到推論運算的環境使用

Caffe2 Chainer Microsoft Cognitive Toolkit mxnet TensorFlow PYTORCH theano

人工智慧運算平台軟體

IS/IDY Simple to Smart

R 750xa XE8545

前端 推論運算 (INFERENCE)

數據中心透過網路進行判斷

GRE + TensorRT 推論運算最佳化效能軟體

Tesla T4

嵌入式 結合機器人設備進行判斷

JETPACK SDK 嵌入式系統開發軟體

Jetson TX2

無人自動化駕駛車

DriveWorks SDK 無人自駕車開發軟體

Drive PX2

NVIDIA 研發 輔助AI運算所需軟體

cuDNN 與人工智慧運算平台軟體整合的程式庫

NCCL GPU晶片之間運算最佳化溝通軟體

cuBLAS GPU晶片最佳化數學矩陣運算軟體

cuSPARSE GPU晶片最佳化大型稀疏矩陣運算軟體

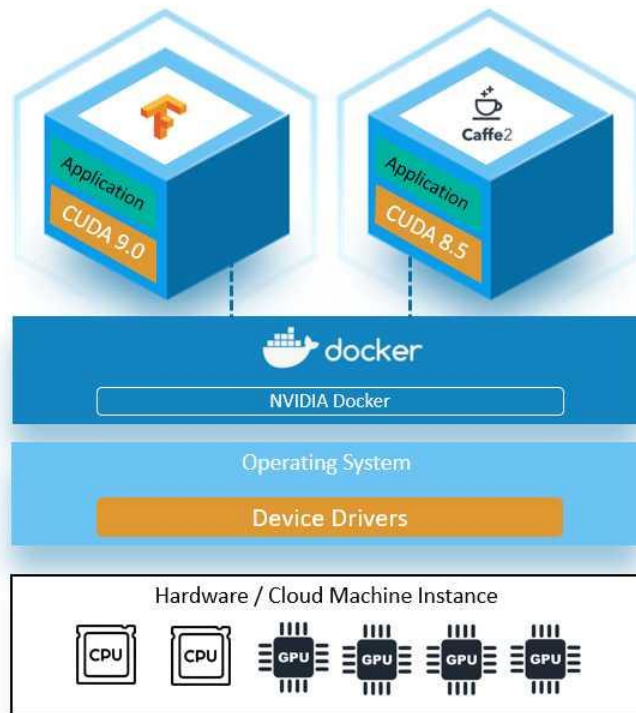
TensorRT 人工智慧推論運算效能最佳化運算軟體



吳宣儀 博士

數位無限軟體股份有限公司營運長

AI基礎設施的趨勢：GPU + 容器



AI應用框架

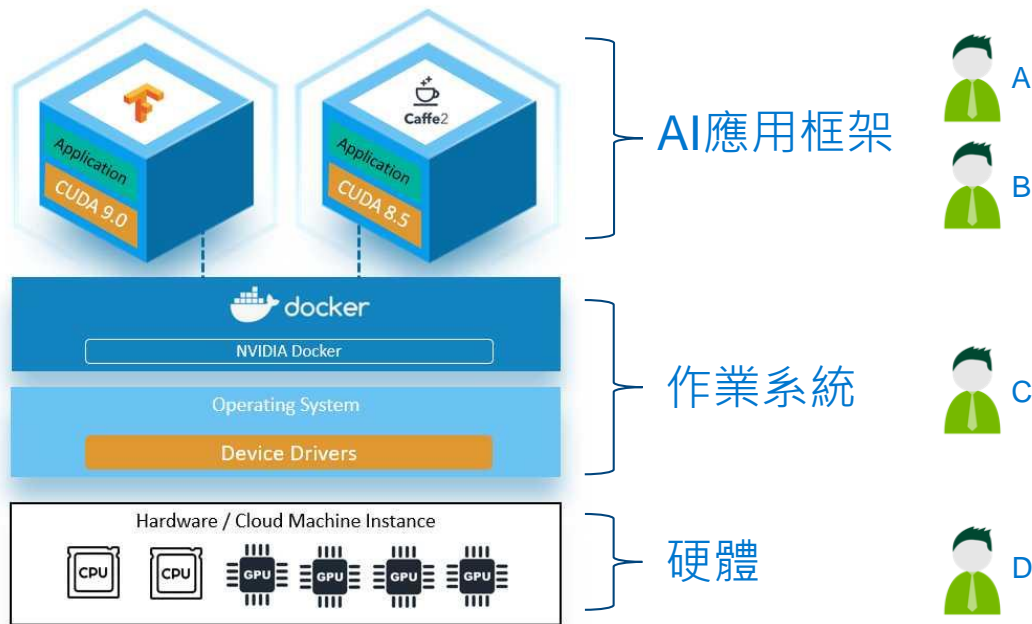
作業系統

硬體

客戶的痛點 !!!



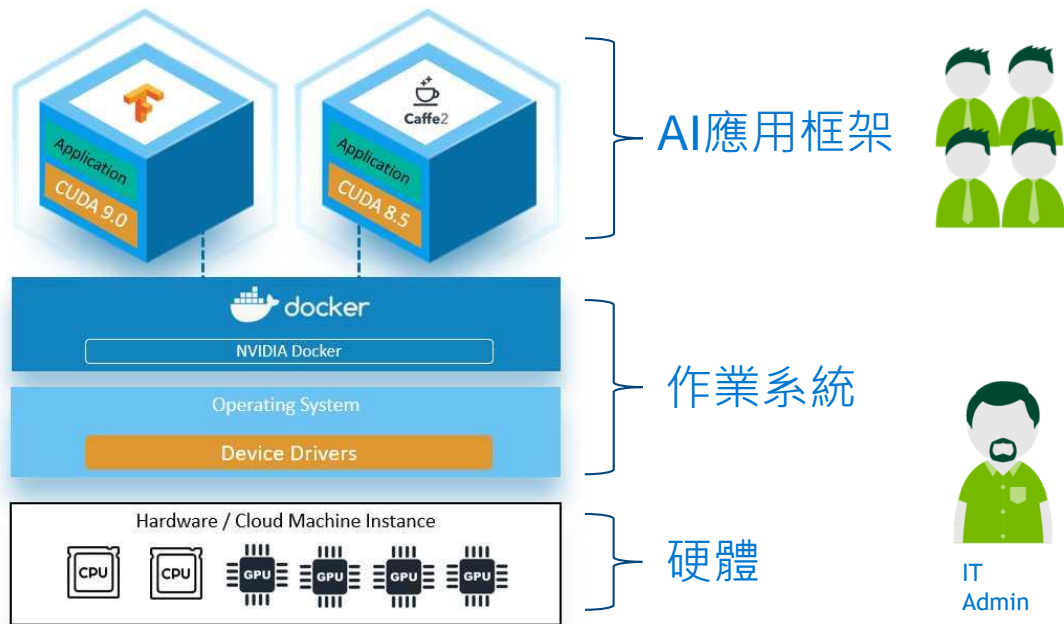
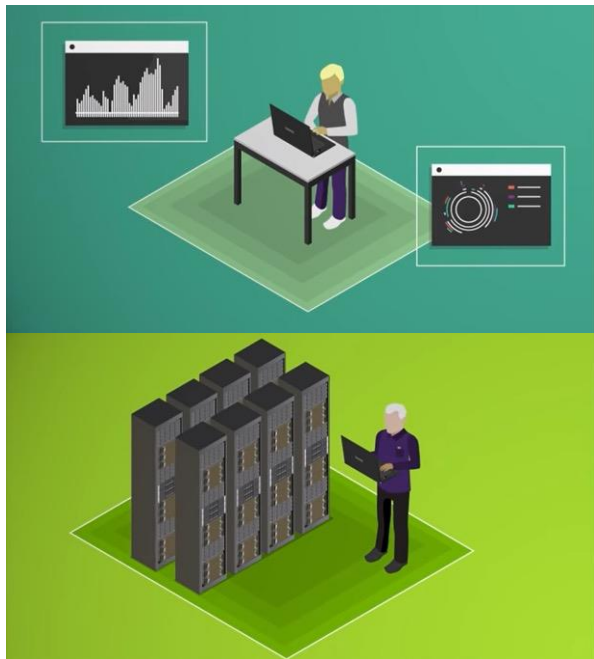
AI 研發人員自用自管



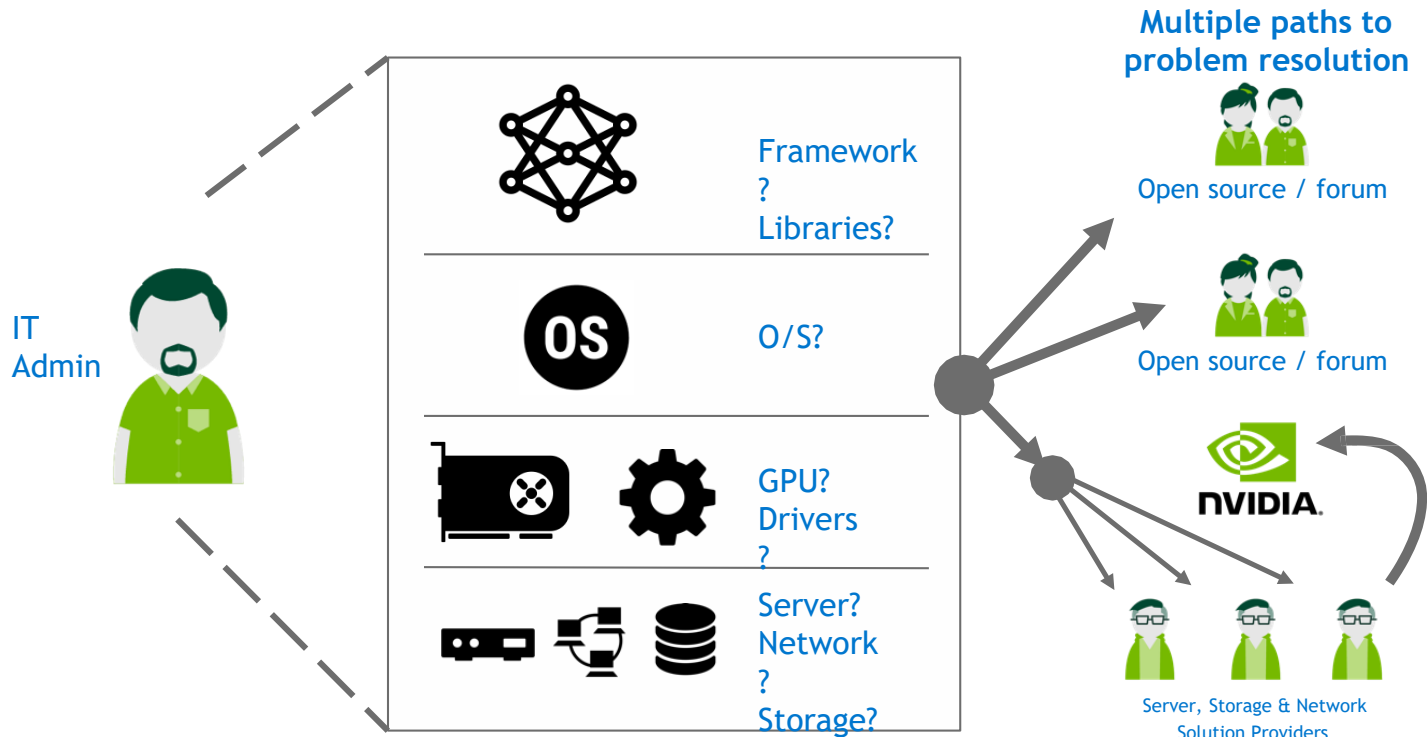
常見問題

- GPU資源無法隔離，占用到他人資源卻不自知
- 需另外學習docker操作指令、自行管理軟硬體環境
- 底層架構權限暴露在外，一人下錯指令，全機無法使用
- 無身分認證，濫用占用也無法追查，只能等待資源釋放
- 各做各的AI應用框架，重複下載浪費資源

IT人員加入管理



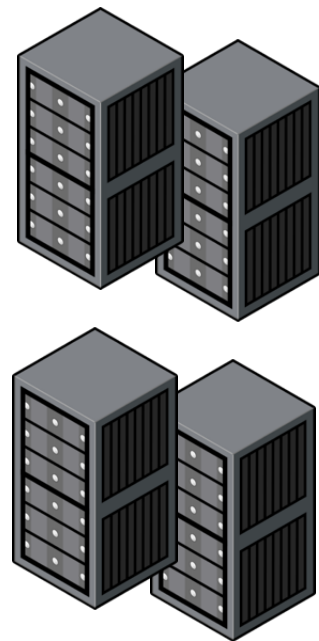
IT 人員需要確認並處理眾多 RD 人員的需求



RD 人員需要 IT 人員提供資源



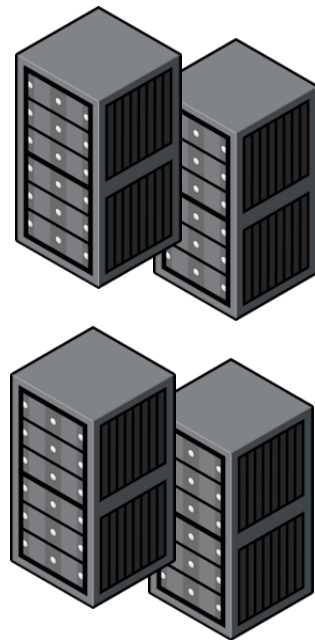
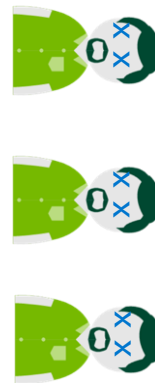
IT
Admin



RD 人員成長需求增加 IT 人員忙不過來



IT Admin



常見問題

- 使用者需求繁複，管理雜亂無章
- 需求又急又趕，打亂工作排程
- 主管關切資源使用狀況，彙整資料不易
- 資源掌握度差，誰再用、用多久、剩餘資源皆不透明
- 需具備docker與kubernetes架構經驗



InfinitiesSoft AI-Stack on Dell

最佳化您的 AI 管理平台

AI-Stack on Dell

RD



IT Admin



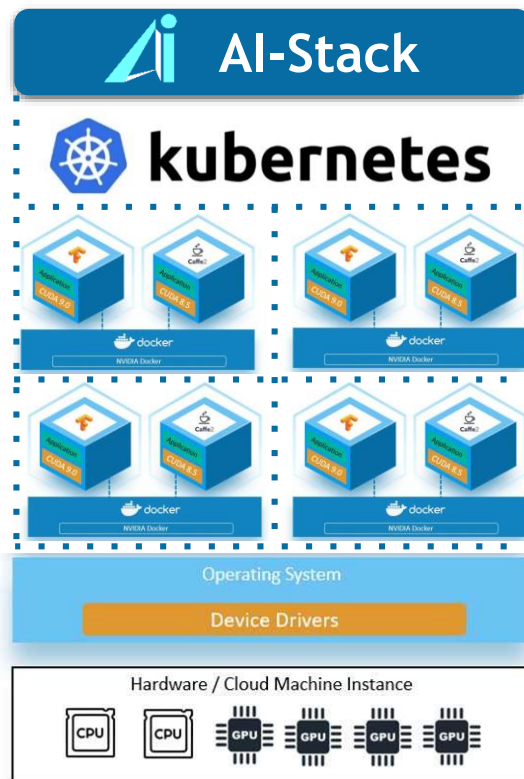
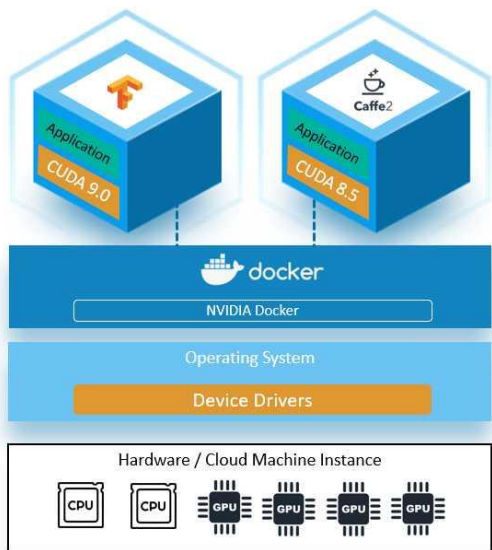
Infinitesoft
AI-Stack

DELLTechnologies

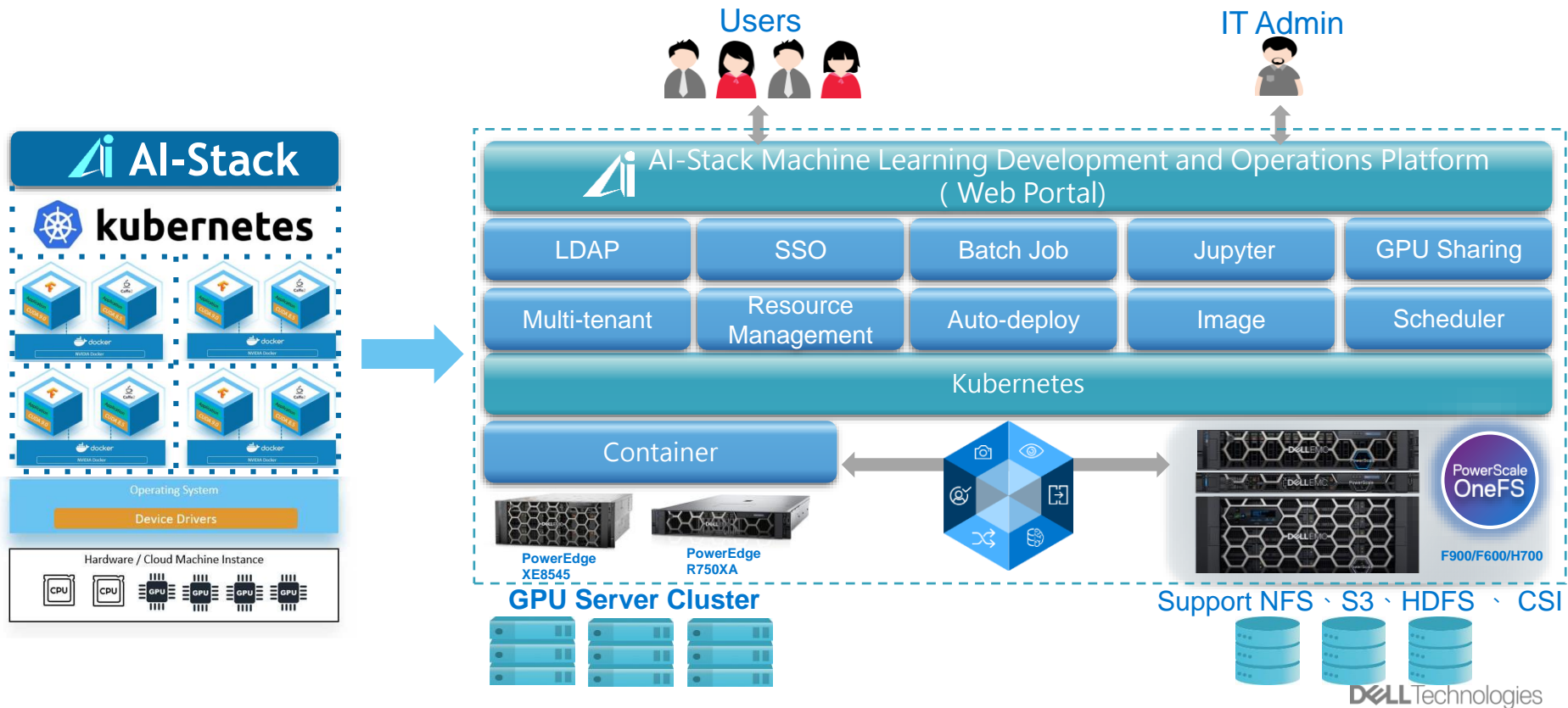
“Is there any easy and fast way to use/maintain GPUs and Storage?”

*“ Yes! **AI-Stack on Dell** is your best solution!”*

What is AI-Stack?



AI-Stack on Dell



AI-Stack on Dell 滿足用戶需求

降低使用者門檻

簡易步驟快速建立 (個人/團隊) ML環境

掌握有效資源

彈性IT資源共享、額度限制、工作排程

環境自主權

IT環境隔離、SSH Key或密碼登入環境

易於使用和共用

帳號/儲存整合，批次、預約申請作業

彈性靈活儲存管理

空間管理，放大，複製，快照，備份還原，異地備援

快速建立機器學習框架與運算資源

← INFINITIES) 機器學習服務) 建立容器

Group B yuntech

機器學習服務

基本資訊 1 建立容器 2

建立容器

容器列表

自定義鏡像

機器學習任務

網路安全服務

帳號管理

Version 3.18.2

GPU 型號

P100 (016GB) V100 (032GB) V100-L (032GB) V100-NV (032GB)

各節點目前可用 P100 (16GB) 片數如下。若欲使用高於 2 片 GPU 的容器可能需要等候建立

2 片

GPU 片數

1 片

全域 GPU 使用剩餘度: 15 + 1 / 18 片

硬體配置

24 CORE + 120 GB RAM

個人 CPU 使用剩餘度: 0 + 24 / = Core

個人 RAM 使用剩餘度: 0 + 120 / = GB

共享記憶體

啟用

鏡像類型

公共鏡像 自定義鏡像

Framework

tensorflow 20.02-tf2-py3

SSH 密碼

再次確認密碼

管理金鑰

建立新金鑰

tylest hankyun

選用白名單

允許所有連入流量

批次建立 啟用

配置費用 TWD 0 / 小時

上一步 下一步 | 建立

← INFINITIES) 機器學習服務) 容器列表

Group B yuntech

機器學習服務

建立容器

容器列表

自定義鏡像

機器學習任務

網路安全服務

帳號管理

Version 3.18.2

搜尋

+ 建立自定義鏡像

名稱	GPU 配置	Framework	部署 ID	建立時間	擁有人	狀態
isstest1229	1 P100 (016GB)	tensorflow:20.02-tf2-py3	8472fb06-a221-492b-8d...	2020/12/29 13:31:57	iss	運行中

部署詳細資訊 服務資訊 監控

外部 IP 10.110.10.10

SSH

```
ssh -i hankyun.pem root@10.110.10.10 -p 31770
```

Jupyter

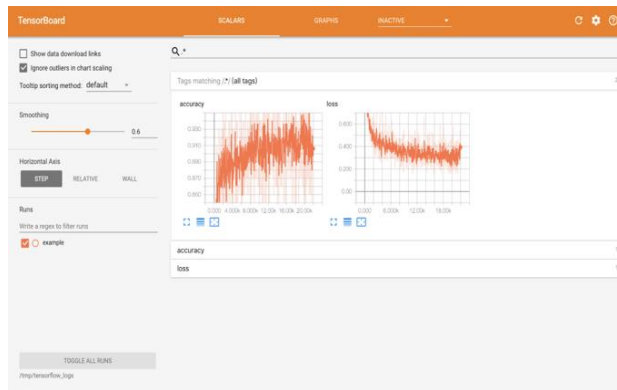
連結

JupyterLab

連結

無縫銜接AI/ML研究員常用的開發工具

```
Wed Jan 30 04:00:02 2019
=====
HWID2A-D8C 395.30
=====
GPU Name Persistence Mode-10 0 100%
GPU Temp Perf PowerUsageCap MemoryUsage Utilization Computer Age
-----
0 Tesla V100-DMC... 0C 100000000.00 0 0
N/A 37C 70 440 / 3004 11818 / 16383632 04 Default
=====
Processes:
GPU PID Type Process name Usage
-----
No running processes found
=====
In [1]: from _future_ import print_function
import tensorflow as tf
# Import MNIST data
from tensorflow.examples.tutorials.mnist import input_data
mnist = input_data.read_data_files("./mnist/train-images-idx3-ubyte.gz", one_hot=True)
# Parameters
learning_rate = 0.01
training_epochs = 10
batch_size = 100
display_step = 1
log_dir = "/tmp/tensorflow_logs/example"
# If Graph input
# mnist data image of shape 28x28x3
```



Web-based terminal interface for a container. The top bar shows 'INFINITIES | 機器學習服務 | 資源列表'. The main area displays system information and a terminal session. A text box on the right says: '透過「網絡終端機」可直接在 Web UI 連線至容器進行操作'.

```
INFINITIES | 機器學習服務 | 資源列表
-----
Password:
Last login: Fri Jul 26 18:52:09 UTC 2022 on pts/1
welcome to ubuntu 18.04.3 LTS (GNU/Linux 4.15.0-112-generic x86_64)

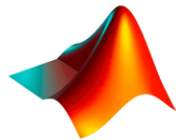
Documentation: https://help.ubuntu.com
Management: https://landscape.canonical.com
Support: https://ubuntu.com/support

This system has been minimized by removing packages and content that are
not required on a system that users do not log into.

To restore this content, you can run the 'unminimize' command.
ubuntu@futures1320-67409096c-1a161-5:~$ ls -l -li
total 44
drwxr-xr-x 1 webuser webuser 4096 Jul 26 18:53 .
drwxr-xr-x 1 webuser webuser 4096 Jul 26 18:49 ..
-rw-r--r-- 1 webuser webuser 127 Jul 26 18:54 .bash_history
-rw-r--r-- 1 webuser webuser 3773 Jul 26 18:49 .bashrc
drwxr-xr-x 2 webuser webuser 4096 Jul 26 18:50 .cache
-rw-r--r-- 1 webuser webuser 867 Jul 26 18:49 .profile
drwxr-xr-x 1 webuser webuser 4096 Jul 26 18:49 .ssh
-rw-r--r-- 1 webuser webuser 4096 Jul 26 18:52 .vim
-rw-r--r-- 1 webuser webuser 1238 Jul 26 18:53 .viminfo
drwxr-xr-x 1 root root 2048 Jul 26 20:29 docker-examples
drwxr-xr-x 2 root root 4096 Jan 30 20:29 docker-examples
ubuntu@futures1320-67409096c-1a161-5:~$
```

Screenshot of the MATLAB R2020b software interface. The window title is 'MATLAB R2020b - academic use'. The interface shows the 'HOME' tab with various toolbars for file operations, workspace management, and code execution. The 'Command Window' is open, displaying the message: 'How to MATLAB! See resources for Getting Started'.

無縫銜接AI/ML研究員常用的開發工具



MATLAB®



TensorFlow



TensorBoard



PyTorch



Keras

Caffe



Caffe2

mxnet



Cognitive Toolkit

GPU分配利用狀況一目了然

← Operation) 機器學習服務) GPU 節點

機器學習服務

- MLS 規格
- MLS 樣板
- 節點資訊
- GPU 節點**
- 規格
- 事件紀錄
- 錯誤處理
- 使用者
- 系統

Version 3.17.3

GPU 節點資訊

供應商 **IBM** 區域 **us-east-1** 查詢

總覽

Used 1.5 of 4

- 節點 2 台
- 容器 15 台
- 資源 GPU * 4

節點

0.5 0.3 未分配

Used 0.8 of 2

- 節點 **gpu1**
- 資源 2 * V100 (32 GB)
- IP 位址 140.11.10.10
- 容器 8 台
- 維護 設置

容器資訊

進階檢視

租戶名稱	容器名稱	擁有人	GPU 占用比
afy	matlab1	charles	10 %
afy	matlab2	charles	10 %
afy	matlab3	charles	10 %
afy	matlab4	charles	10 %
class2	pytorchdev	ai-sec3	10 %

0.7 未分配

- 節點 **gpu2**
- 資源 2 * P40 (24 GB)
- IP 位址 140.11.10.11
- 容器 7 台
- 維護 設置

← INFINITIES) 帳號管理) 成本分析

機器學習服務

機器學習任務

網路安全服務

帳號管理

- 雲平台設置
- 操作紀錄
- 基本資料
- 訂單紀錄
- 審核管理
- 成本分析**
- 個人化設定

以月為週期 < 2020-09-01 ~ 2020-09-30

服務類型 **MLS**

已花費用 \$65,846.20 TWD

上週期已花費用 \$126,725.80 TWD

預估花費 \$66,484.20 TWD

平均每日花費 \$3,873.31 TWD

金額(TWD)

時間(日)

Kubernetes

市面上唯一支持MIG技術的AI平台

← Operation (機器學習服務) GPU 節點

GPU 節點頁面增加「多執行個體 GPU」專區，呈現已啟用 MIG 的節點

多執行個體 GPU

MIG 資源配置中... 重新整理

節點 test-mig... IP ... 10.20.1...
資源 A100 (4... 容器 0 台
MIG 配置中

點按「設置」可對該節點上支援 MIG 的 GPU 卡進行切割或還原

MIG 配置

GPU List

0	已切斷	20GB	20GB
1	已切斷	10GB	10GB
2	已切斷	10GB	10GB
3	已切斷	10GB	N/A
4	已切斷	5GB	5GB
5	已切斷	5GB	5GB
6	已切斷	5GB	5GB
7	已切斷	5GB	N/A

取消配置 清除選項 暫存配置

取消 確認送出

輕鬆掛載儲存設備

INFINITIES 機器學習服務 儲存管理 IT Team region-Kubem...

機器學習服務 搜尋

建立容器 容器列表 自定義鏡像 GPU 使用率 儲存管理

名稱	擁有人	儲存種類	權限	狀態
<input checked="" type="checkbox"/> ineedvolume1	mock	nfs-storage1	僅個人可讀寫	可用
<input type="checkbox"/> ineedvolume1-1	mock	nfs-storage1	僅個人可讀寫	可用
<input type="checkbox"/> ineedvolume2	mock	nfs-storage2	租戶可讀寫	可用
<input type="checkbox"/> ineedvolume3	mock	nfs-storage3	租戶可讀	可用

詳細資訊

僅個人可讀寫
 租戶可讀寫
 租戶可讀

名稱	擁有人	儲存種類名稱	狀態	容量
ineedvolume1	mock	nfs-storage1	可用	31415 GB

INFINITIES 機器學習服務 儲存管理 IT Team region-Kubem...

機器學習服務 更新 Storage

建立容器 容器列表 自定義鏡像 GPU 使用率 儲存管理

名稱* ineedvolume1

權限* 僅個人可讀寫
 租戶可讀寫
 租戶可讀

儲存種類* nfs-storage1

已掛載容器 容器名稱: cont01 擁有人: happygo(我自己)

更新 *無法複製已掛載的裝置權限。權限對未來新增裝置生效

INFINITIES 機器學習服務 建立容器 IT Team region-Kubem...

機器學習服務 建立容器

掛載儲存裝置

名稱	擁有人	儲存種類	權限
<input checked="" type="checkbox"/> ineedvolume1	mock	nfs-storage1	僅個人可讀寫
<input type="checkbox"/> ineedvolume2	mock	nfs-storage2	租戶可讀寫
<input type="checkbox"/> ineedvolume3	mock	nfs-storage3	租戶可讀

名稱	擁有人	儲存種類	權限	容量	掛載路徑*	預設為 Jupyter 工作目錄
ineedvolume1	mock	nfs-storage1	僅個人可讀寫	31415 GB	/mnt/a01	<input checked="" type="checkbox"/>

取消 確認

上一步 下一步 | 建立 儲存裝置

INFINITIES 機器學習服務 建立容器 IT Team region-Kubem...

機器學習服務 建立容器

轉換類型* 公共鏡像 自定義鏡像

Framework* testcreatecustomtemplate01

SSH 密碼*

應用白名單 允許所有進入流量

掛載儲存裝置 啟用

儲存裝置 選擇儲存裝置

儲存裝置名稱	掛載路徑	權限	預設為 Jupyter 工作目錄
ineedvolume1	/mnt/a01	僅個人可讀寫	<input checked="" type="checkbox"/>

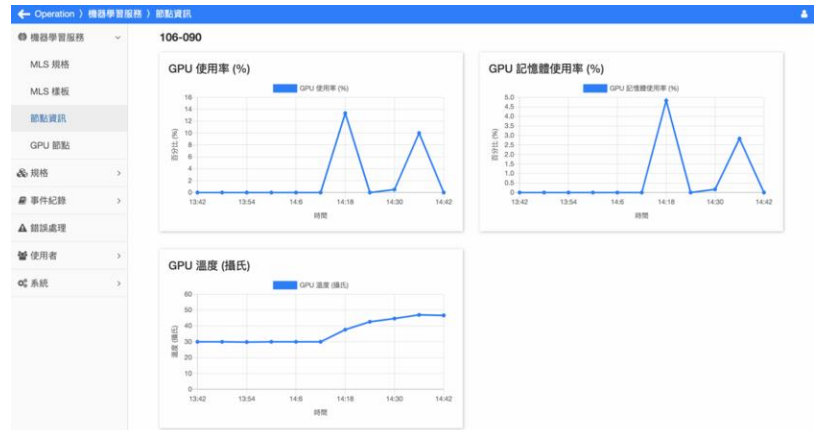
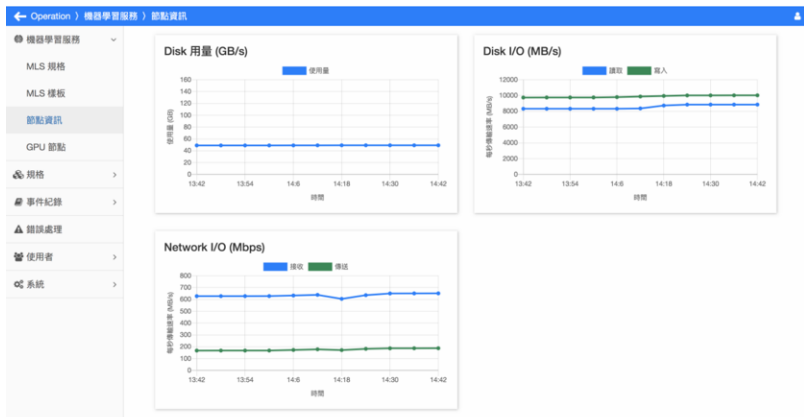
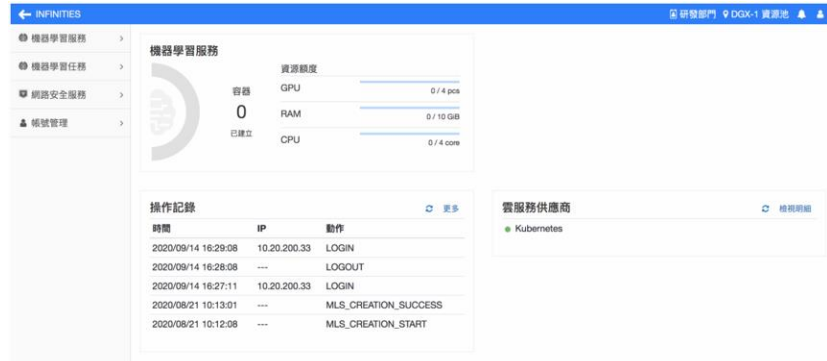
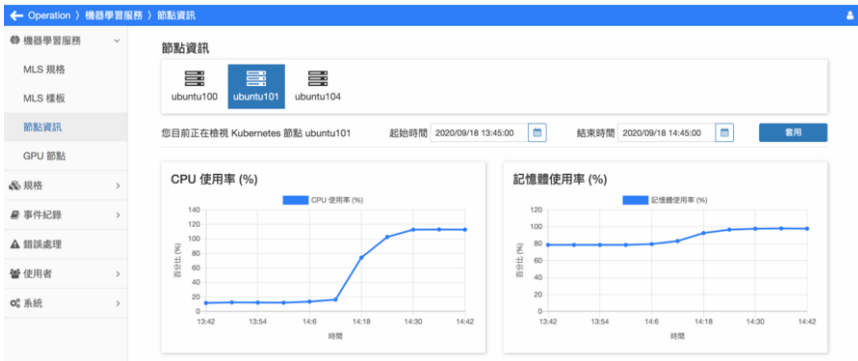
批次建立 啟用

標籤的儲存裝置中擁有權限為個人權限者，不得啟用批次建立

配置費用 100 / 小時

上一步 下一步 | 建立 清除選擇

各項資源監控



Dell | AI-Stack 滿足強大易用的需求



Dell + AI-Stack 優勢

- 資源高效分配，透過用戶帳號認證快速將GPU算力交到使用者手上
- 強化用戶資源高度自助使用，降低管理者日常操作負擔
- 精細掌控資源使用成本與報表功能，管理昂貴資源不費力
- 內建資源使用審核功能，讓管理者全方位掌握資源流向
- 掛載外部儲存與GPU伺服器擴容彈性
- 計費計價、成本分析機制
- GPU共享，更有效的提升GPU利用率
- 業界最佳擴展及靈活Dell PowerScale 儲存設備、快速的空間部署及資料管理

關於 InfinitiesSoft

- 亞洲最早開發的混合雲管理平台
- 台灣第一個GPU雲的開發者
- AI-Stack為台灣市佔率最高的AI PaaS



DELL Technologies



重點指標客戶

科研機構



政府機關



Metro Taipei



半導體



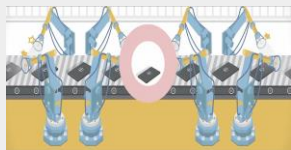
基建事業



Dell Technologies + AI Stack 解決方案全景圖

AI場景化應用落地解決方案

應用層



AI+製造
產品檢測, 良率提升, 預防維護



AI+零售
商品識別, 智能稱重, 智能貨櫃



AI+教育
人工智慧教學實訓平臺



AI+金融
智慧語音, RPA



AI+醫療
影像輔助診斷

面向AI應用就緒的平臺解決方案

平臺層

Ready solution for AI(Nvidia)

資源調度, 雲化, 視覺化



VMware Bitfusion



面向GPU的AI就緒解決方案

GPU 虛擬化彈性計算架構

Orion



GPU 動態加速雲

Ready solution for AI(Intel)



Intel AI 協議棧的就緒解決方案



深度學習框架



計算資源管理與優化庫

GPU

vGPU

FPGA

CPU 分散式訓練

Graphcore IPU

AI-Stack

端到端AI基礎架構解決方案

基礎架構層



AI 計算伺服器
GPU/FPGA/IPU



GPU 工作站



Isilon
ME4084



資料交換網路



DPS 資料保護



VxRail (HCI)



HPC 就緒解決方案

DELL Technologies

總結

以數據及AI為核心，驅動企業創新

- 以數據+AI整合平台、一站購足、附以靈活的付費模式
- 善用雲原生自動化部署策略～加速企業創新 Time-to-Market
- 邊際、核心、雲端場景都能部署～企業能因地制宜，開展具靈活性及彈性
- 員工行動力提升～企業抗疫生存能力提升

DELL Technologies **Can Help!**

DELLTechnologies