

以異質性融合基礎架構加速催化人工智慧醫療發展

Catalyzing Acceleration in Medical AI with Heterogeneous Infrastructure

雲達科技

張文祥

Stephen Chang

2024/11/05





Agenda

現代異質性融合基礎架構在加速智慧醫療發展的角色



優化基礎架構的擴展性與靈活性以加速智慧醫療發展



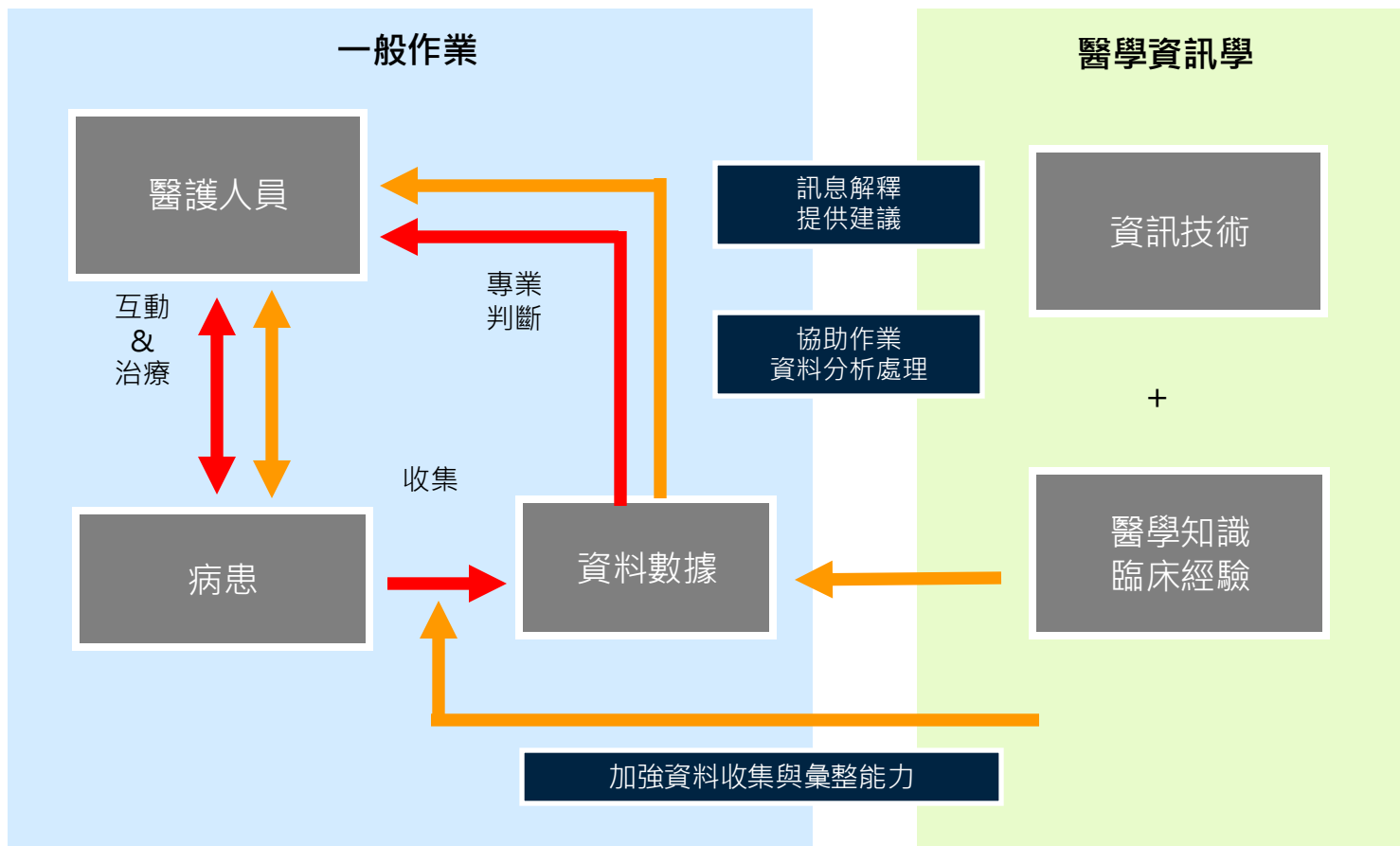
透過異質性融合基礎架構來實現智慧醫療



現代異質性融合基礎架構在加速智慧醫療發展的角色



將電腦資料處理技術結合通訊技術應用於醫學知識與臨床經驗並將所得的資料做最快速且合理的推論與歸納，並發展一套最佳運作模式或系統，以提升醫護服務的效率與準確性且減少資源浪費



依需求醫療人力與專業知識的介入程度，將資訊技術於醫療應用概分為六個層級：

- 資料擷取與傳輸
- 資料儲存與讀取
- 資料分析處理
- 診斷與決策建議
- 治療與控制
- 研究與發展



Health/Fitness

Medical

Health Care

健康管理

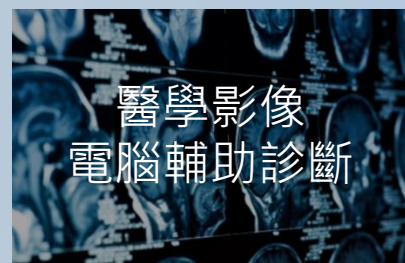
疾病預防

疾病診斷

疾病治療

癒後/復健

...



...

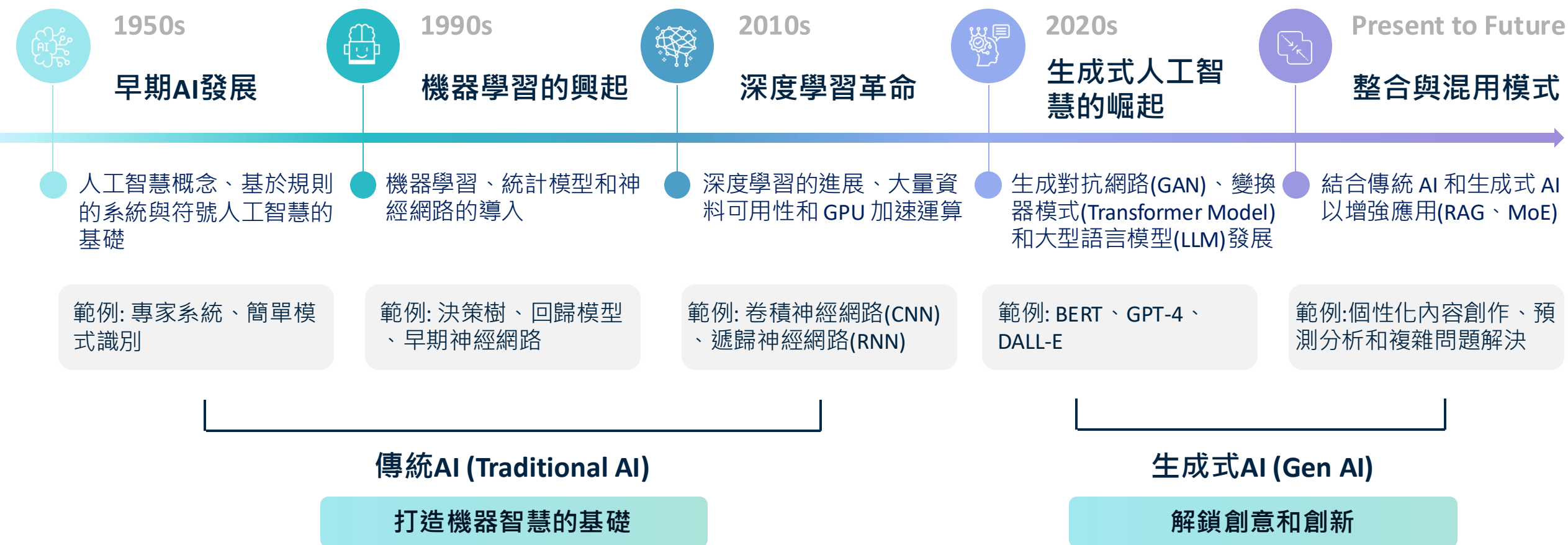
AI Training / Inference / HPC :人工智慧模型訓練與應用、高速運算

Big Data / Data Analytics : 巨量資料儲存、整合、分析

Edge/ IOT: 資訊收集、傳輸

資料來源：工研院產科國際所





使用AI工具來自動化重複性任務以減化流程及負擔、從大量數據中提取有價值的資訊，或者利用機器學習模型來預測未來的趨勢或生成資訊，或者是建立智能系統來自動化複雜的決策流程



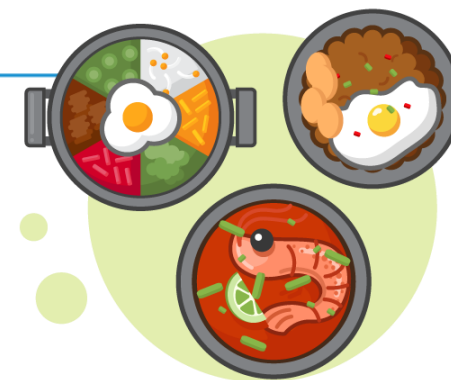
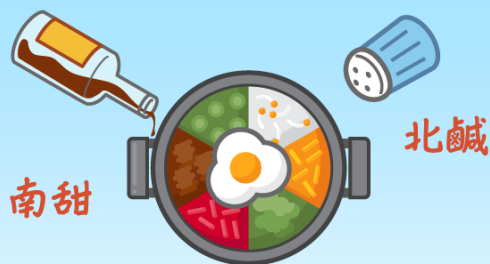
將AI結合終端應用，協助各產業改進工作流程、處理問題，
或實現更好的結果，以實踐數位轉型

AI學習 (AI Training)

AI應用 (AI Inference)

預訓練 (Pre-training)

微調訓練 (Fine-tuning)



雲端機房



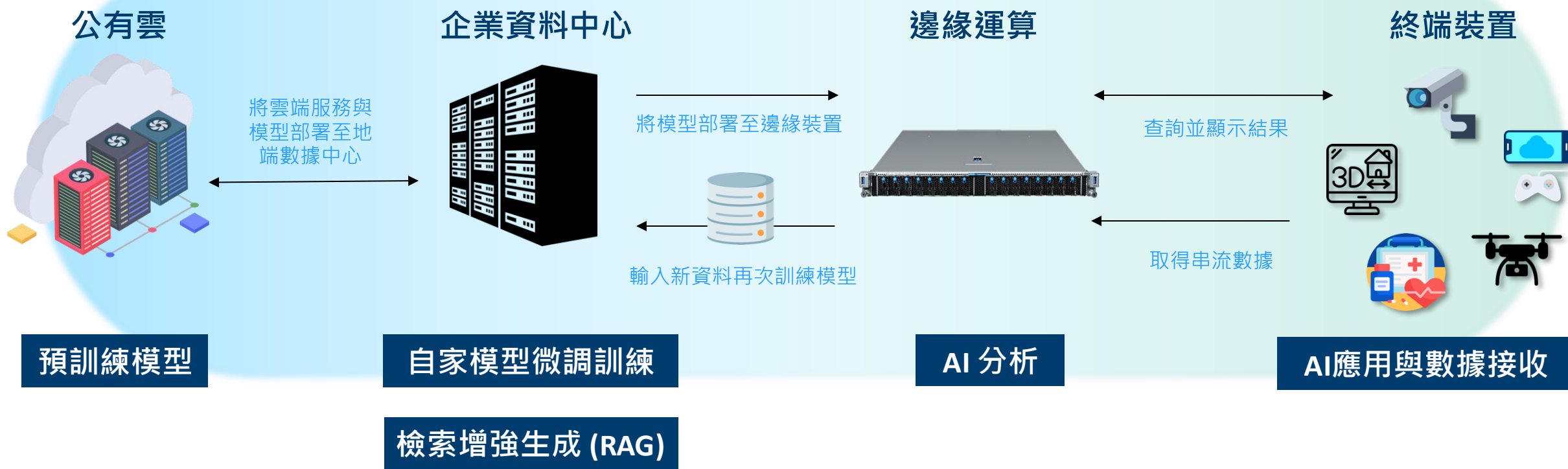
邊緣

雲端

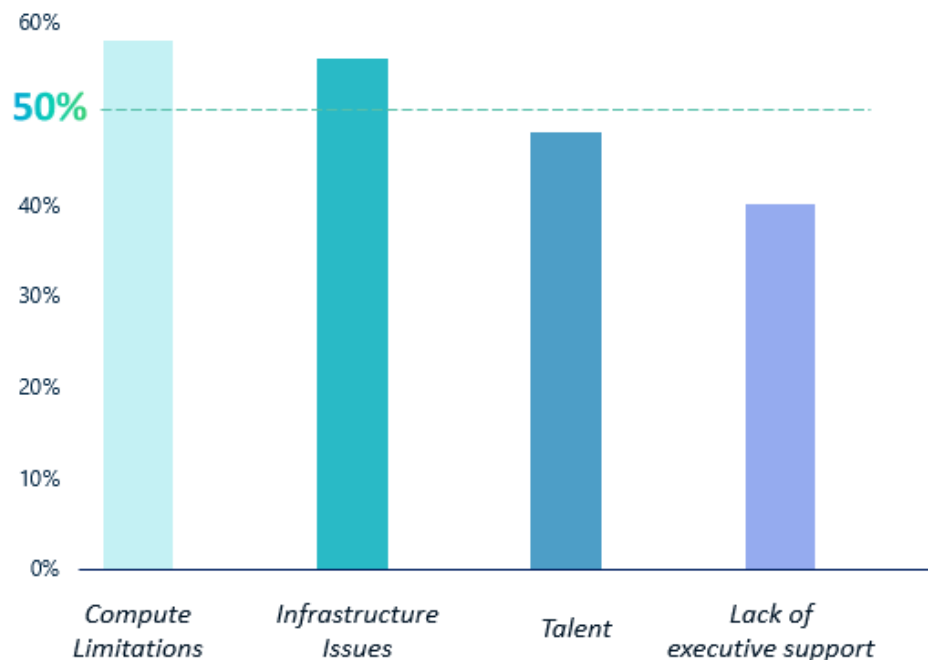
強大的運算力與數據支援AI模型訓練

邊緣

低延遲支援即時性AI推論分析



Respondents (%)



Source: The State of AI Infrastructure at Scale 2024

問題:貴組織在擴展 AI 時面臨的最大挑戰是什麼?



超過50%的受訪者將計算限制及基礎設施問題視為 AI 服務部署的主要挑戰

如何打造人工智慧應用資訊基礎架構成為企業在使用AI 服務所面臨的最大課題

傳統AI (Traditional AI)

打造機器智慧的基礎

使用成熟的技術如機器學習、基於規則的系統和專家系統以解決特定的問題



高速運算單一主機

算力著重資料分析推理分析等任務

適度的記憶體及資料儲存資源

稍高的能源消耗

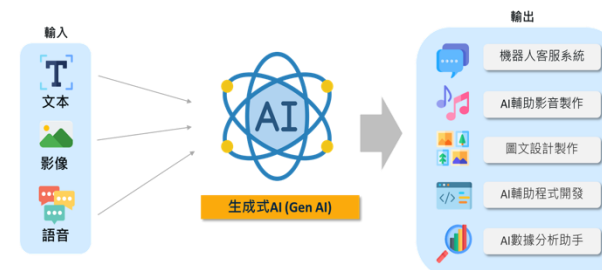
系統架構差異

相較於傳統AI，**生成式AI**對硬體及系統架構有更高的要求

生成式AI (Gen AI)

解鎖創意和創新

讓電腦能夠基於既有的資料，創造出全新且有價值之內容



異質性運算叢集系統

多樣性高算力以執行生成式AI模型應用服務

大量記憶體及儲存資源

分散式高速網路

高能源消耗並搭配高效冷卻系統



製造業



運輸業



醫療業



金融業



高等教育與學研



Numeric Weather Prediction

Computation Fluid Dynamics

Molecular Dynamics

Quantum Chemistry

Electronic Structure

Next Generation Sequencing

Finite Element Analysis

Dynamic Structure Analysis

AI

BI

AR/VR

EDA

SCADA

OLTP / OLAP

CAD / CAM / CAE

Data Lake / DW

HFT / RM

產業應用負載

傳統封閉式單一架構

Bare-metal



HPC workloads



AI workloads



DA workloads

Pain Points

- ✓ 人力資源有限無法同時管理多個獨立系統
- ✓ 資源分配較無效率
- ✓ 成本高且缺乏協同效應

現代異質式融合架構

Bare-metal | Virtual Machine | Container



Converged HPC & AI & DA Workloads

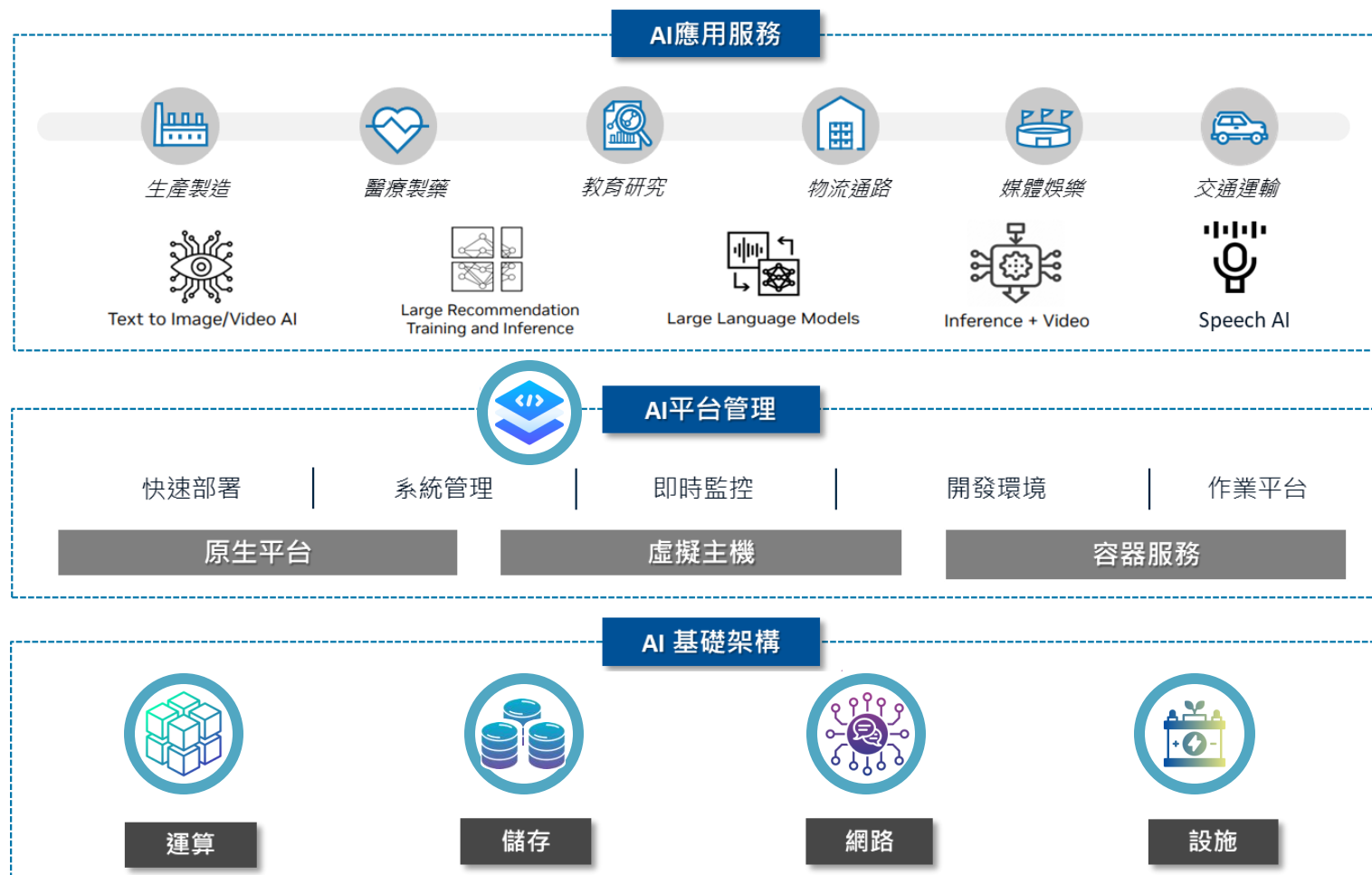
Benefits

- ✓ 方便系統管控
- ✓ 有效處理多種類型應用負載
- ✓ 成本效益高，減少TCO

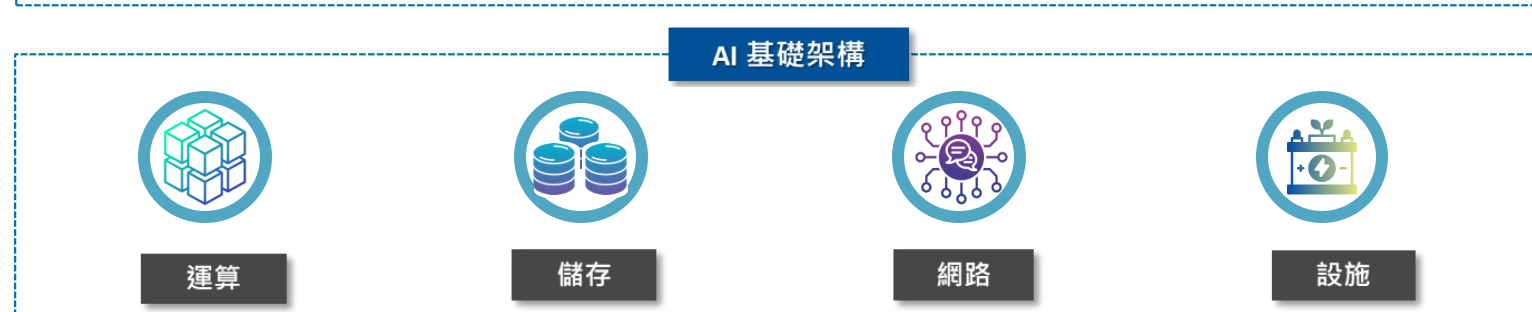
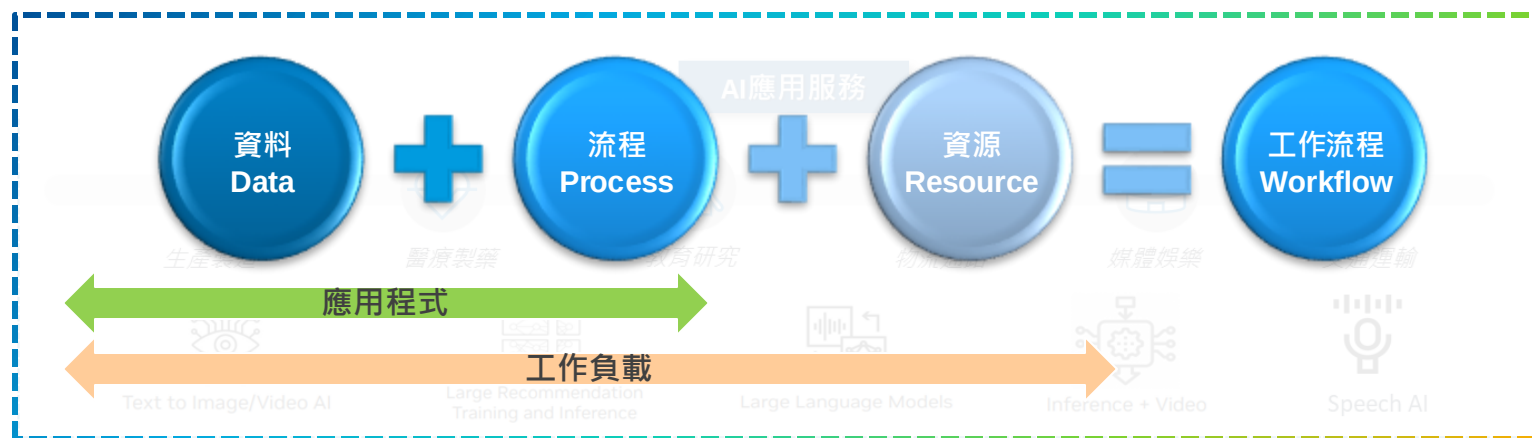
優化基礎架構的擴展性與靈活性以加速智慧醫療發展



根據AI 應用工作負
載的資料工作流程與
特徵以優化系統架構



優化工作負載流程，以發揮系統最大效能、利用率與成本效益



基於工作負載設計的 AI 系統

- 可擴展性 ● 效能瓶頸 ● 異質架構
- 負載平衡 ● 調配策略

最大化AI系統效率

- 以工作負載為中心的策略方法
- 性能優化 ● 資源分配管理 ● 客製化

提高性能，可擴展性&資源利用率

- 性能調校 ● 資源運用 ● 成本效益
- 未來規劃



運算

- 大量的運算力
- 使用大量的GPU處理器或AI加速器
- 採分散式運算架構以支援大型AI模式運算



儲存

- 大規模可擴式資料儲存
- 需要 高效能高容量 資料儲存服務
- 使用SSD及NVMe以提供高速資料存取



網路

- 高頻寬 網路架構
- 低延遲 網路連結
- 主要用於分散式運算與生成資料傳輸



設施

- 先進的機房冷卻方案
- 可擴展式資料中心資訊基礎架構
- 超高耗能及特殊電力配置



軟體

- 具備先進優化特性的軟體框架
- 針對深度學習及生成式AI模式特殊化程式庫
- 專注於分散式運算與雲端服務整合

計算與服務節點

- L組登入服務節點(Login)
- D組部署系統管理節點(Mgmt)
- A組系統服務管理節點(Admin)
- C組虛擬主機/容器服務節點(SVC)
- I組CPU HPC計算節點(CPU)
- J組CPU/BigMem計算節點(BGM)
- K組Cloud Service計算節點(NGS)
- M組GPU/AI Inference節點(GPU INF)
- N組GPU/AI Training節點(GPU TRN)

高性能檔案儲存系統

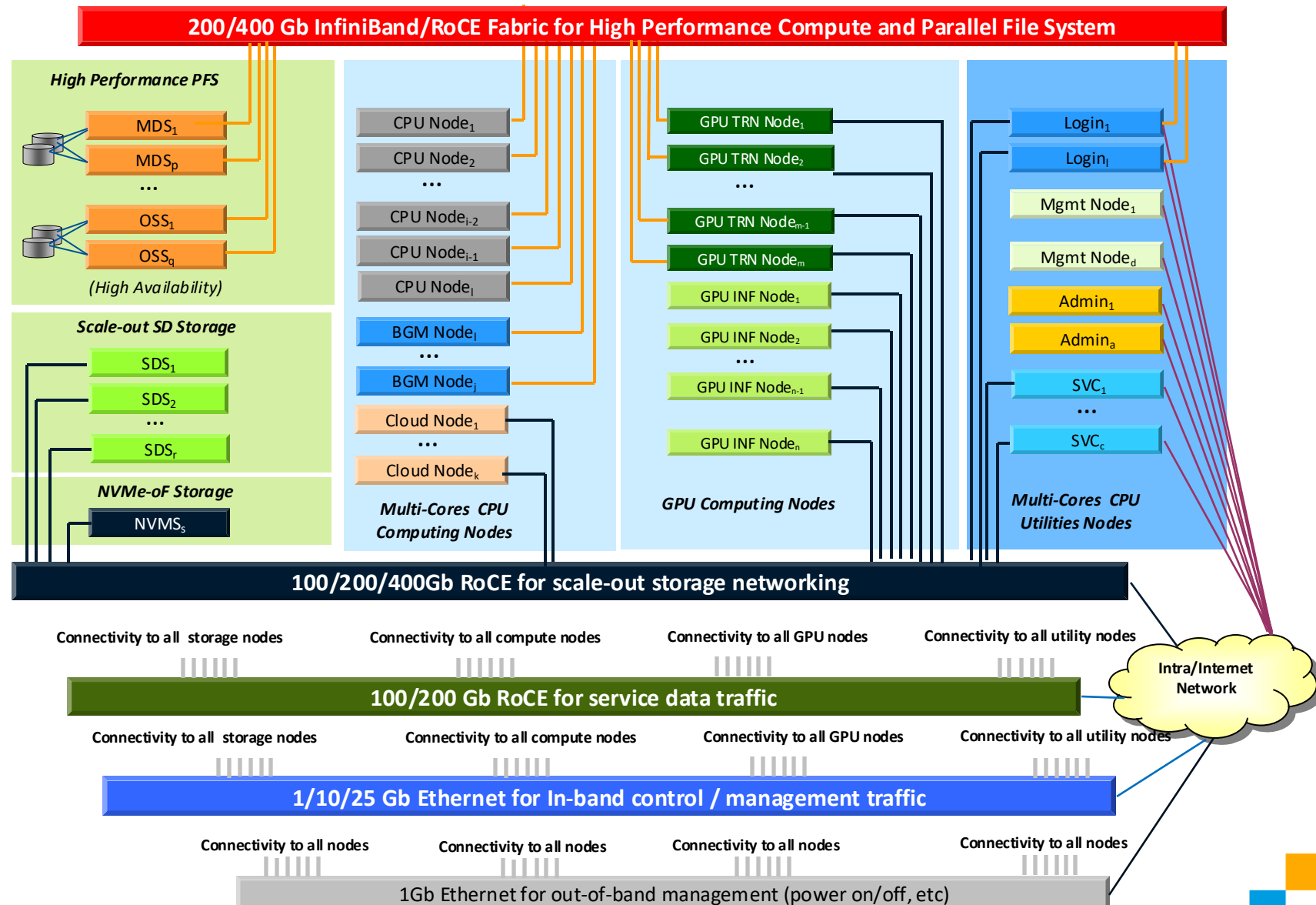
- P/Q組MDS/OSS儲存節點

網路區塊/物件/檔案儲存及備份系統

- R組SDS儲存節點
- S組NVMS儲存節點

計算、管理與資料傳輸網路

- 連接所有節點之1Gb BMC埠至管理網路
- 連接所有節點之10/25GbE 網路埠至控制網路
- 連接所有節點之100/200GbE 網路埠至服務網路
- 連接所有節點之100Gb/200/400Gb網路埠至儲存網路
- 連接計算節點之高速網路埠 (InfiniBand/Ethernet)至運算網路



- 雲端基礎設施

- 提供硬體資源和工作排程管理系統
- 支援裸機(Bare-Metal)、虛擬機(VM)或容器(Container)的工作負載

- 程式開發平台

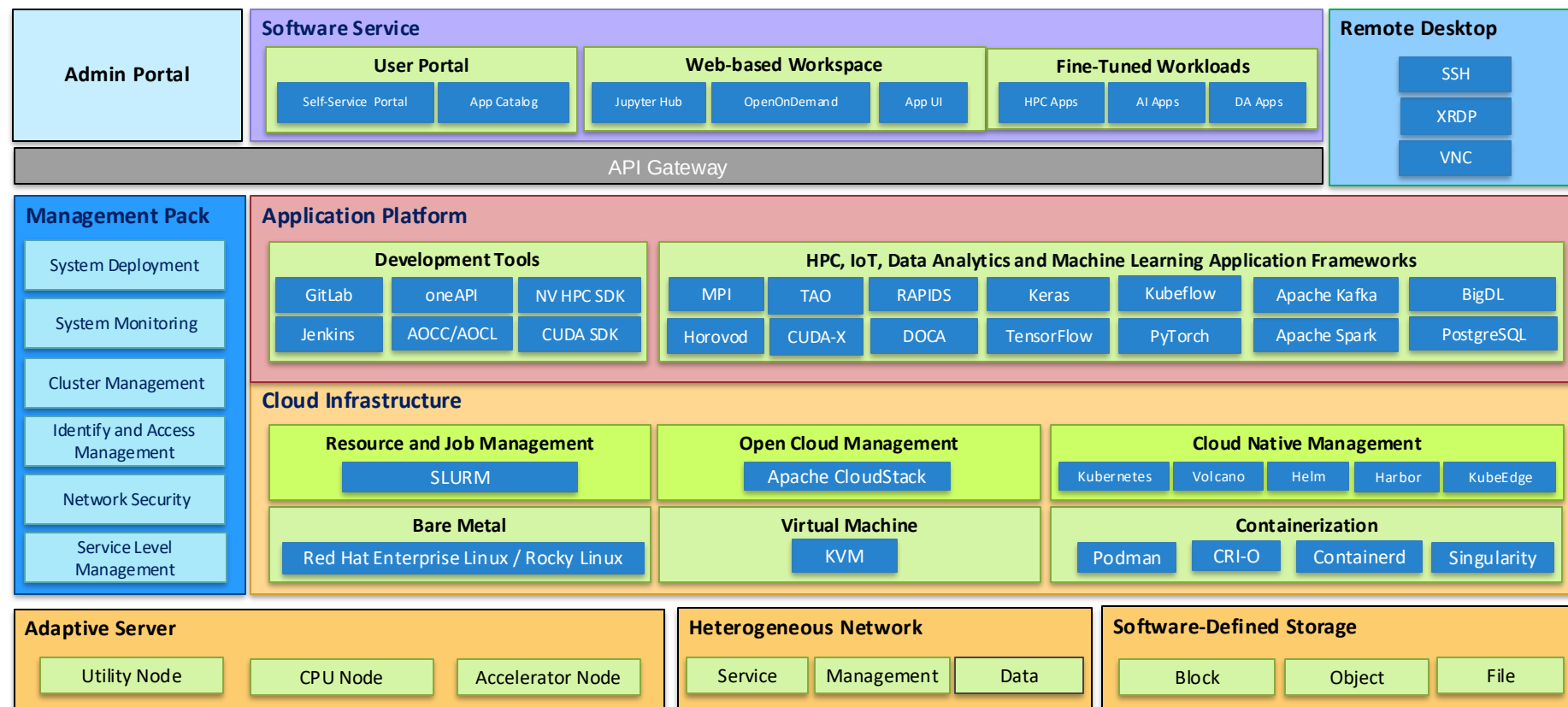
- 包括應用程式開發和管理工具
- 提供編譯器、程式庫、除錯工具及軟體框架

- 應用軟體服務

- 使用者入口網站與網頁式介面操作環境
- 針對機器學習、資料分析、高效能運算和人工智慧工作負載優化
- 整合第三方應用服務

- 系統管理套件

- 用於工作負載性能和狀態管理的工具
- 包含系統部署、監控、管理等工具
- 管理叢集、網路安全和服務等級

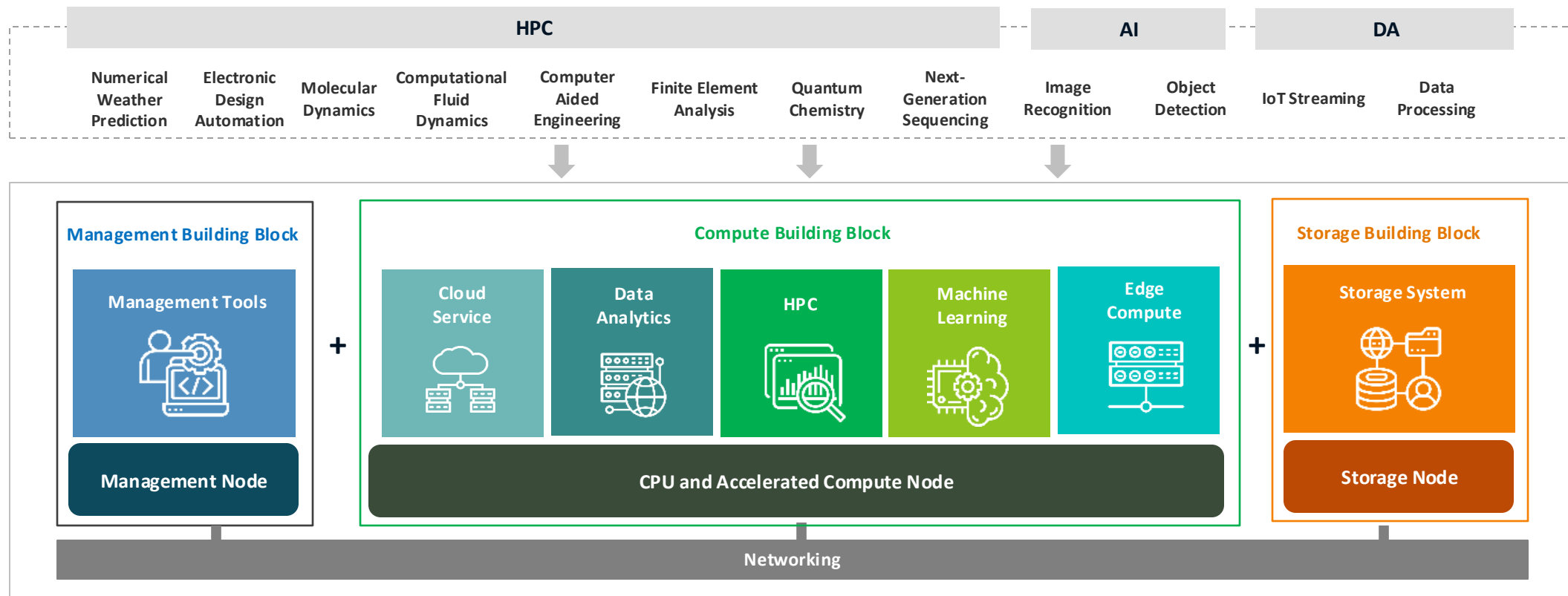


透過異質性融合基礎架構來實現智慧醫療



QCT POD (QCT Platform On Demand)

針對特定工作負載(Workload)以提供最適化軟硬體配置之預先驗證(Pre-Validated) 的 HPC/AI/DA本地(On-Premises)系統



系統管理者

簡化系統部署與管理

- 自動化系統快速部署
- 即時系統監控管理
- 簡化叢集系統管理



開發人員及使用者

加速應用程式開發流程

- 可靠程式開發環境
- 優化應用程式工作負載

資料集管理



ICU MIMIC DB



Dataset

DICOM

Import dicom images from PACS system, and you must have IRB number with hospital...

Strucure Data

Import csv files from DB or your files, and you must have IRB number with hospital. You can...

CSV : 肺癌患者的病歷資料和用藥紀錄

CSV : ICU 患者的基本資訊和生理參數

DICOM : MRI 影像資料集

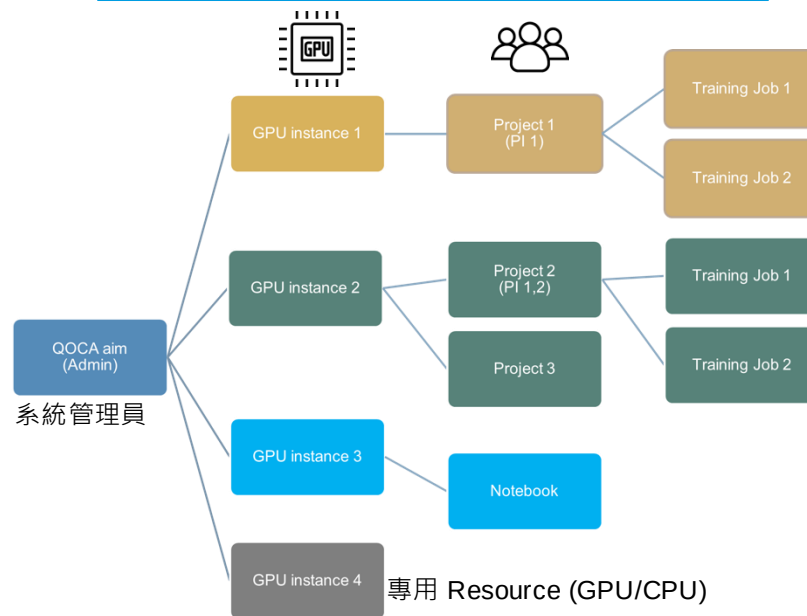
DICOM : CT 影像資料集

MIMIC : 重症患者資料庫

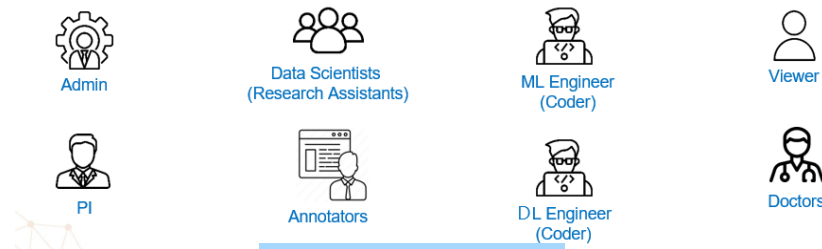
Signal: ECG/EEG 生理訊號圖

管理 AI 所需任何資料

帳號、專案與運算資源管理

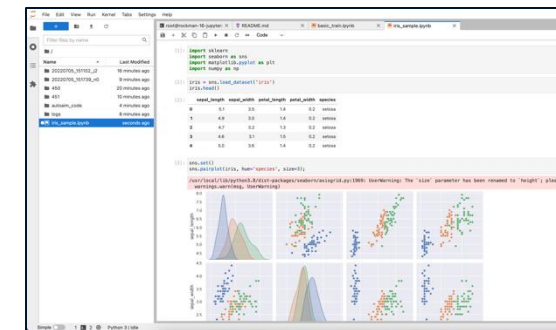
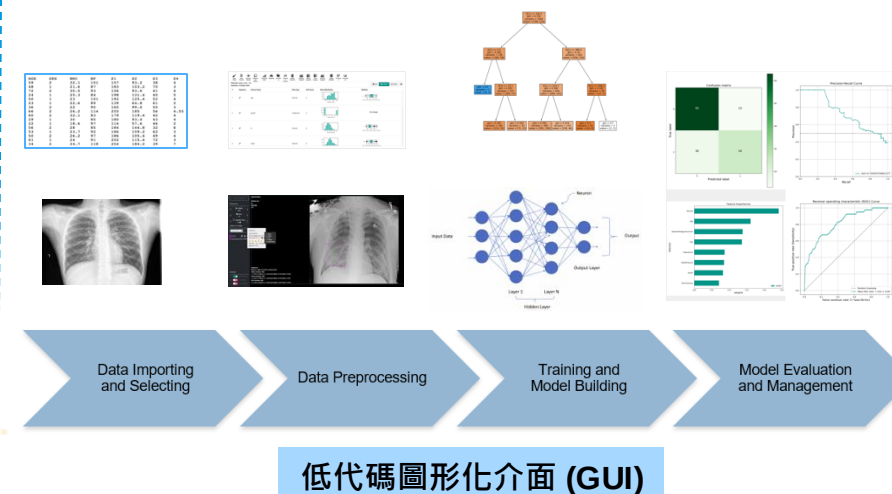


管理系統運算資源

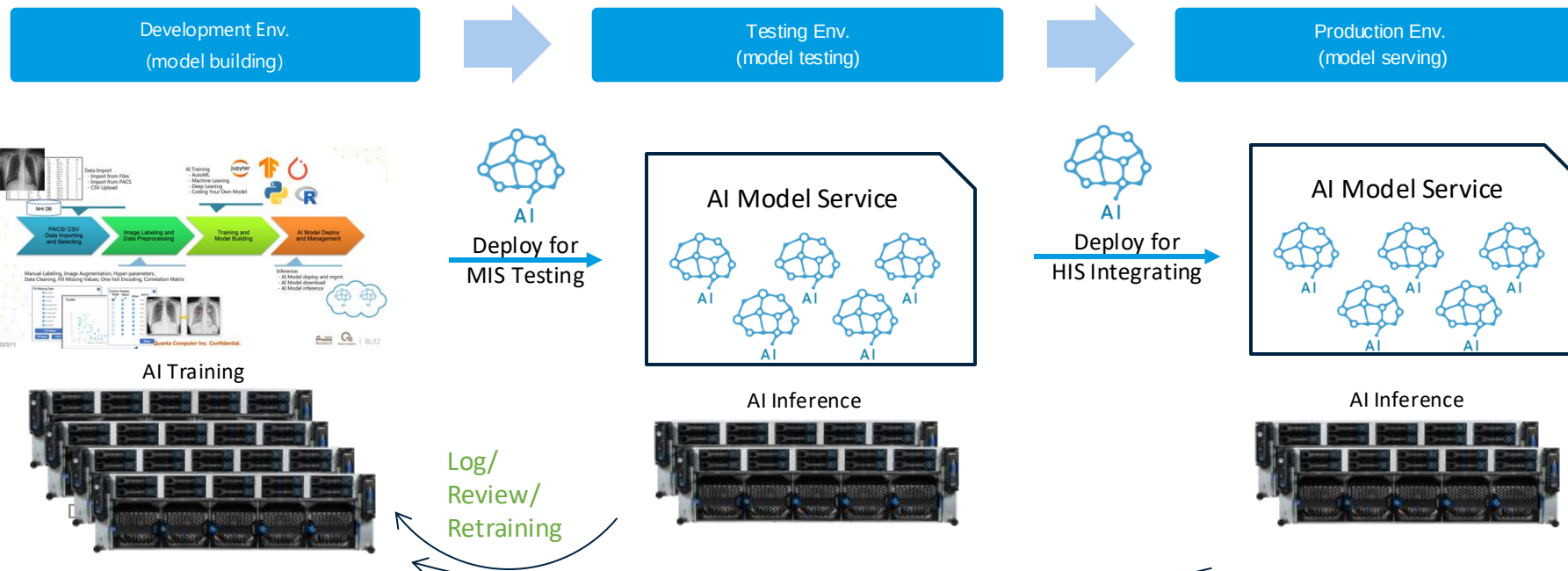
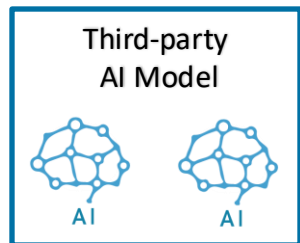
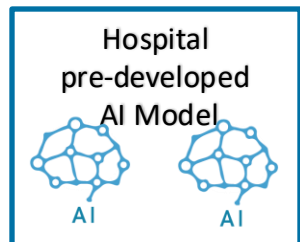
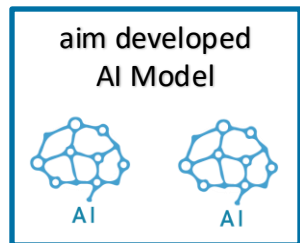


AI 團隊協作與溝通

AI 開發流程與模型管理



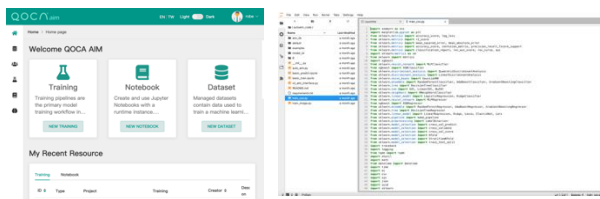
Jupyter Notebook 開發環境



- Clinical and medical dataset
- Structured data, image data
 - MIMIC DB、Time series
 - ECG/EEG、Gene data



Web-based、multi-user
collaborative development



Clinical research, AI educational training,
R&D for HIS/RIS and so on

RESTful API for
AI Model Inference



HIS (OPD, ward, ER), RIS, ...
Hospital testing env.

RESTful API for
AI Model Inference

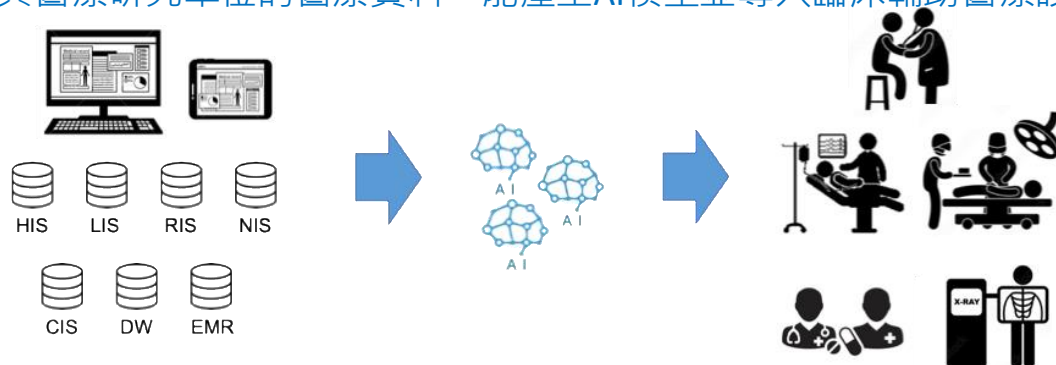


HIS (OPD, ward, ER), RIS, ...
Hospital Production Env.

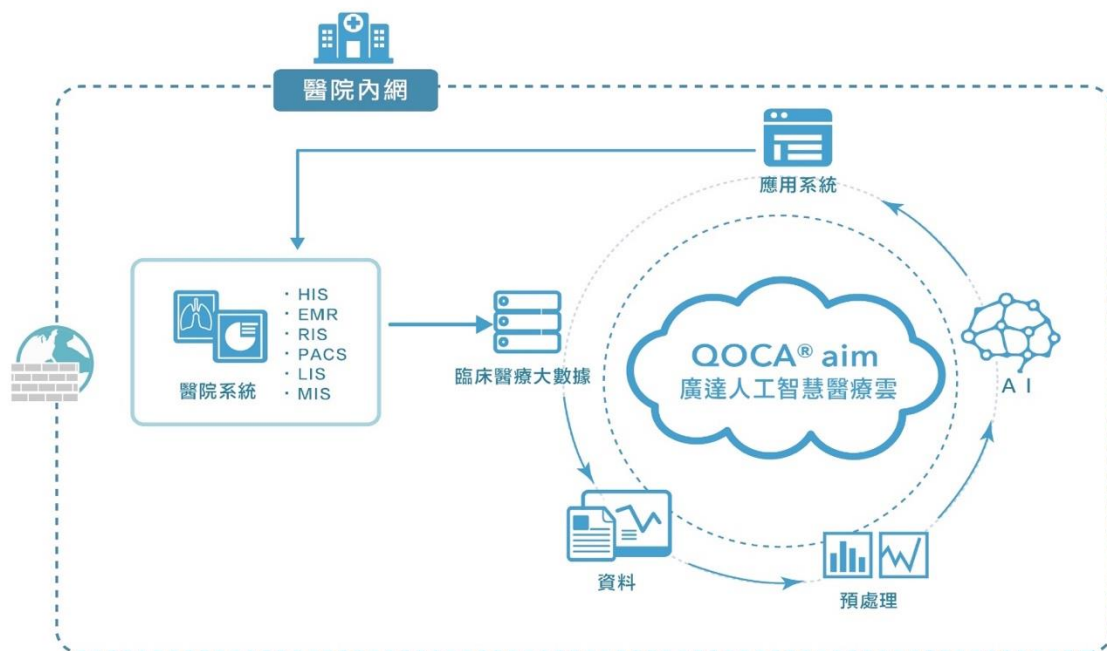
QOCA aim Platform

Data types

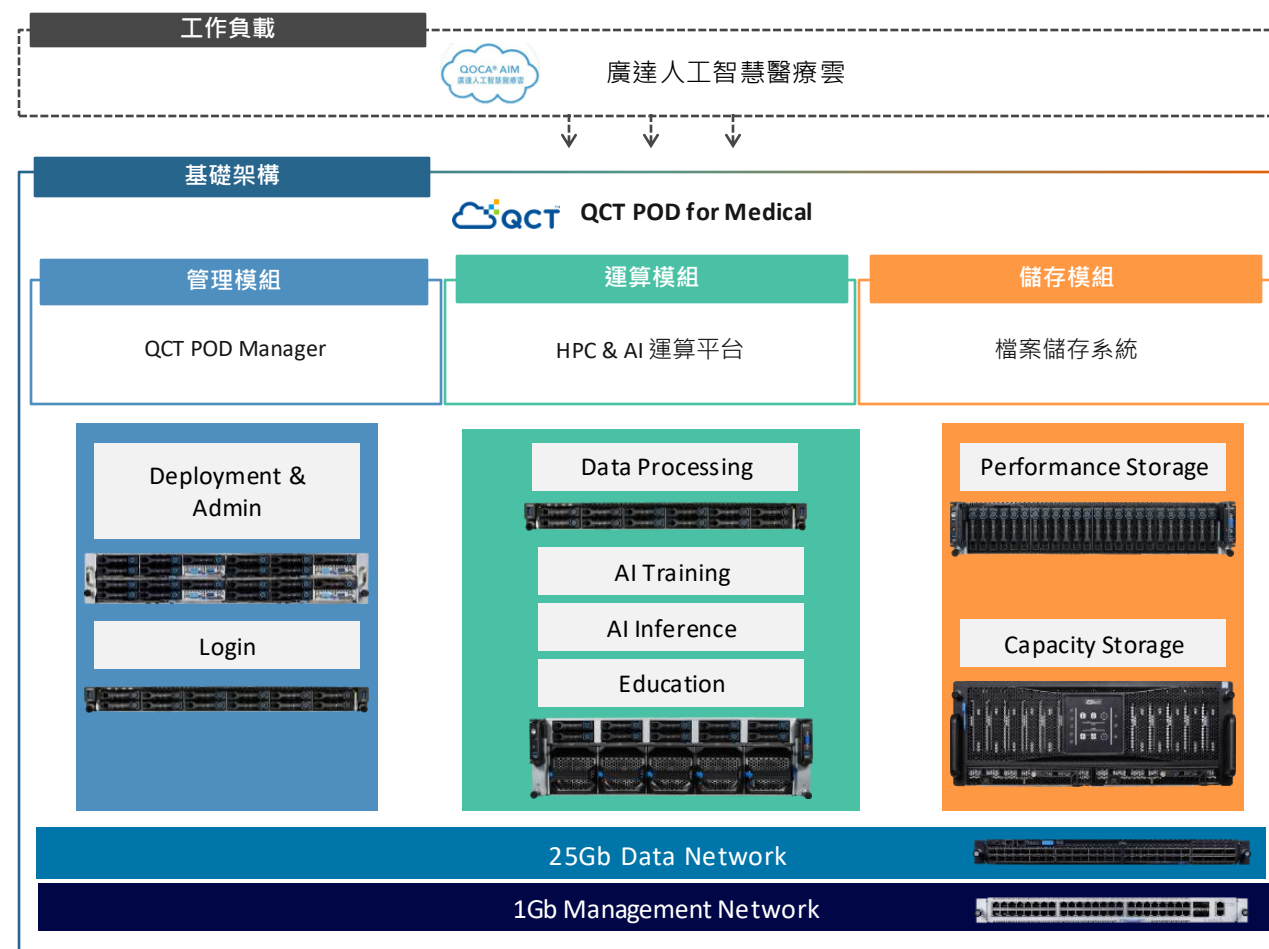
醫院與醫療研究單位的醫療資料，能產生AI模型並導入臨床輔助醫療診斷



高可擴充性系統架構管理硬體資源，建置院內無縫整合醫療資料



QOCA aim - 完整AI整合式開發與應用平台



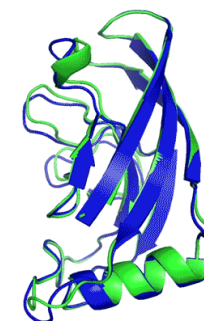
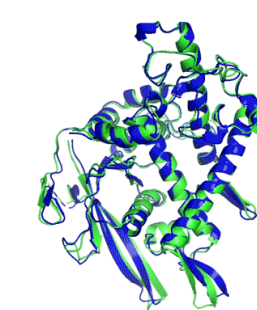
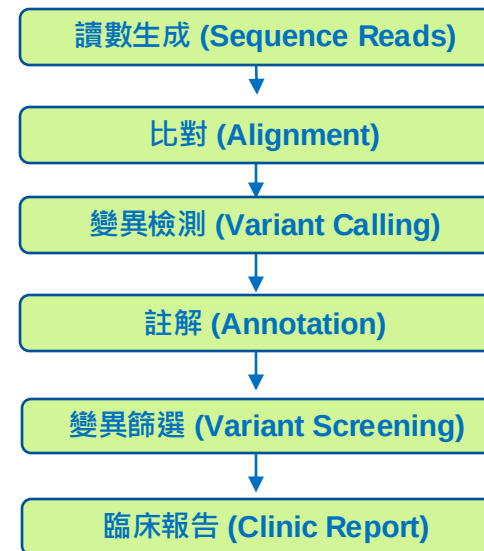
- 何謂AlphaFold2?

- Google DeepMind設計開發
- 一個旨在精確預測蛋白質三維結構的 AI 模型
 - 達到原子級精度(約 1 Å, 10^{-10} m)
- 主要應用領域
 - 藥物發現與設計
 - 疾病機制研究
 - 個人化醫療

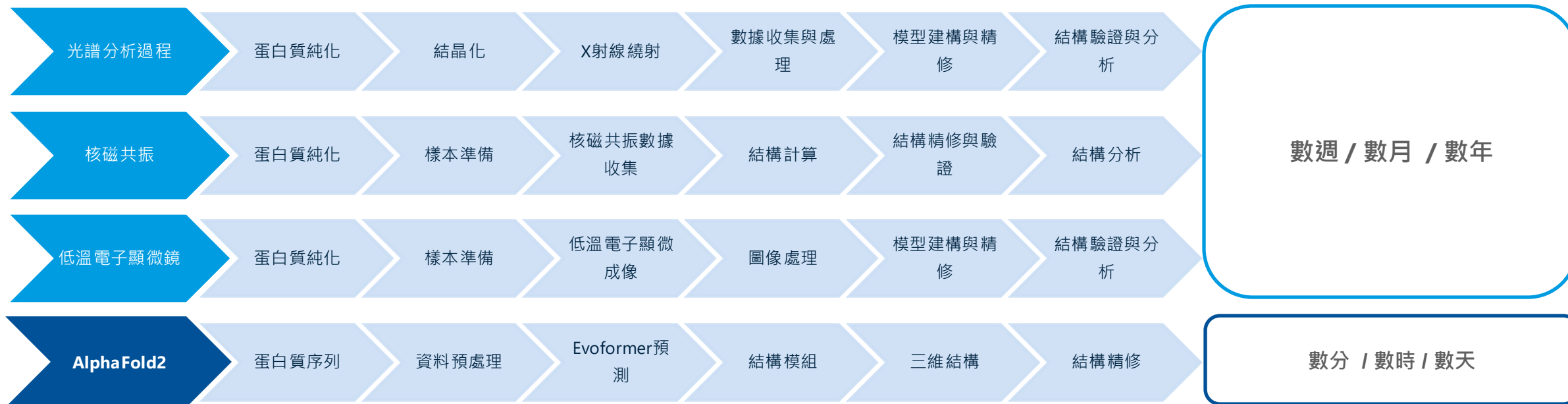
- AlphaFold 2 在次世代基因定序 (NGS) 中的角色

- 促進蛋白質-蛋白質交互作用研究
 - NGS 能快速產生大量基因序列數據，但無法揭示蛋白質的結構。AlphaFold 2 利用氨基酸序列準確預測蛋白質的三維結構，填補了此資訊缺口。
 - NGS 提供序列，AlphaFold 2 產生結構，讓研究者更快了解蛋白質的潛在功能及交互作用部位。
 - 針對 NGS 發現基因變異序列對應的蛋白質結構預測，AlphaFold 2 有助於模擬蛋白質之間的交互作用。這對於構建生物路徑和網路，特別是系統生物學和複雜疾病研究至關重要。
- 加速基因組的功能註解
 - 基因組學的一大挑戰是新發現基因的功能註解。蛋白質結構有助於理解其在細胞中的角色。AlphaFold 2 能預測結構，提供功能域和活性位點等資訊，加快註解過程。
 - 對於非模式生物或缺乏結構數據的蛋白質尤為重要。
- 支持藥物靶點的發現
 - 在藥理基因學和精準醫療中，AlphaFold 2 可預測 NGS 識別出的疾病相關蛋白質結構。這些結構資訊有助於藥物開發和治療靶點識別，優化藥物發現流程。
- 推動個人化醫療
 - 針對個體患者，NGS 可揭示獨特的基因變異。AlphaFold 2 能模擬變異蛋白質，幫助評估突變如何改變結構和功能，為疾病分子機制提供洞見並引導個人化治療。

NGS 分析的核心步驟

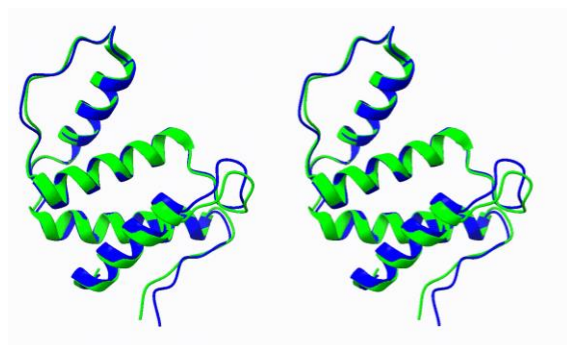


● Experimental result
● Computational prediction



PDB ID: 7FC7
96 amino acid residues

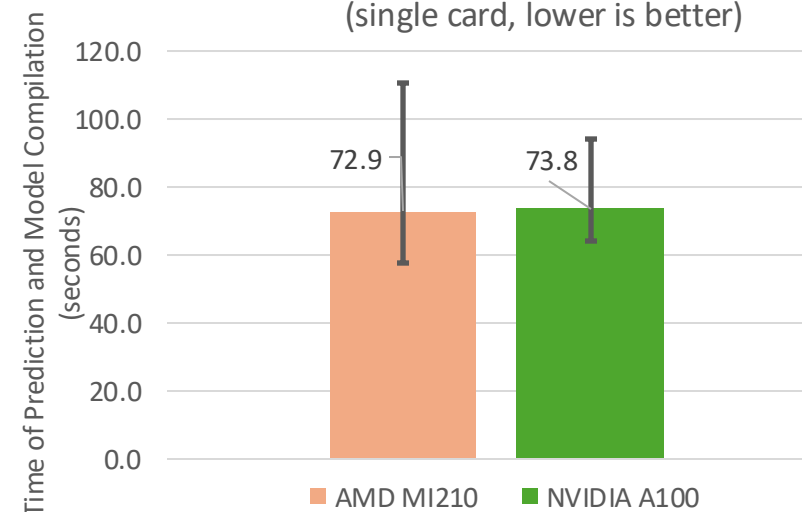
■ 實驗結果
■ 計算預測



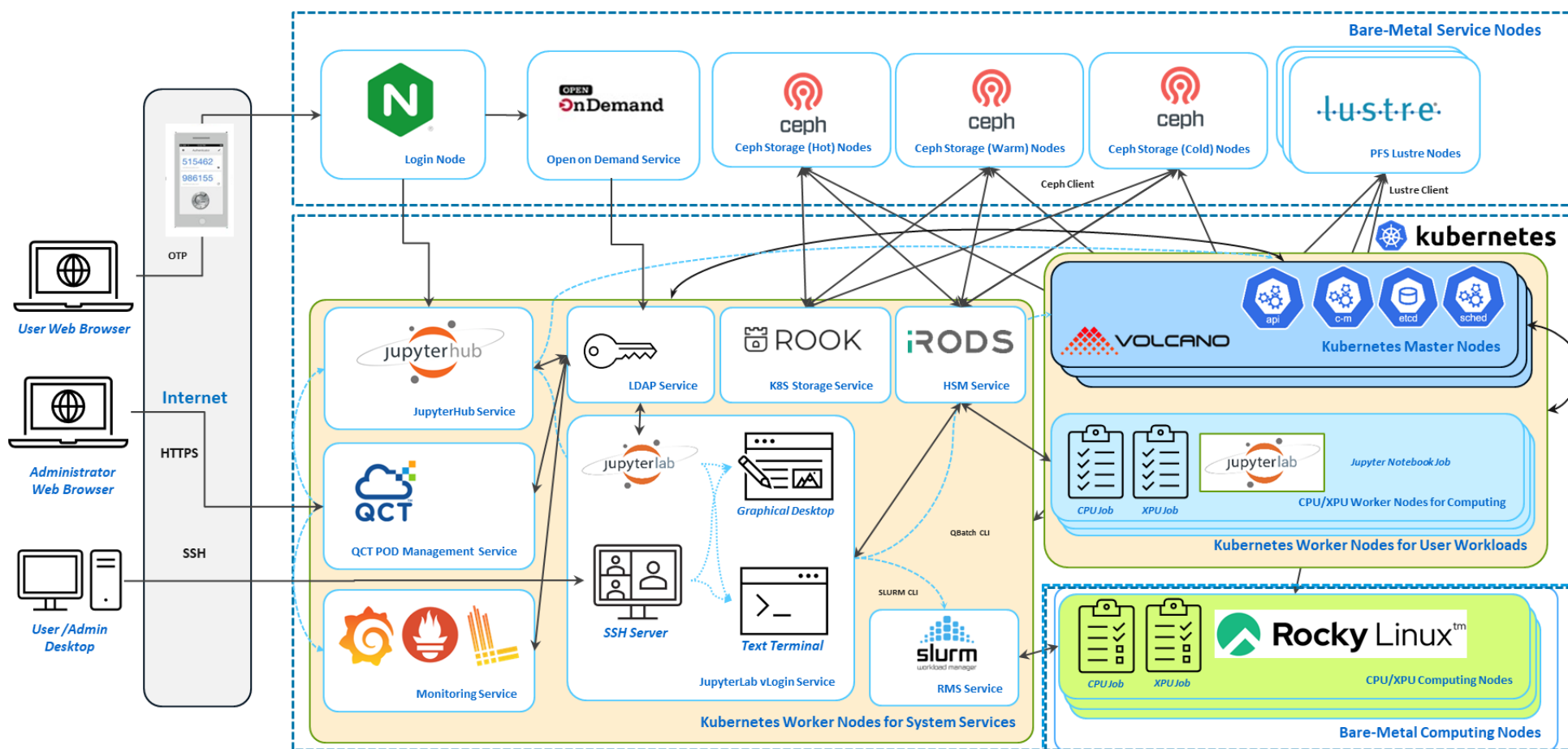
AMD MI210
RMSD: 1.001 Å

NVIDIA A100
RMSD: 1.018 Å

AlphaFold 2 7FC7 Monomer Benchmark
(single card, lower is better)



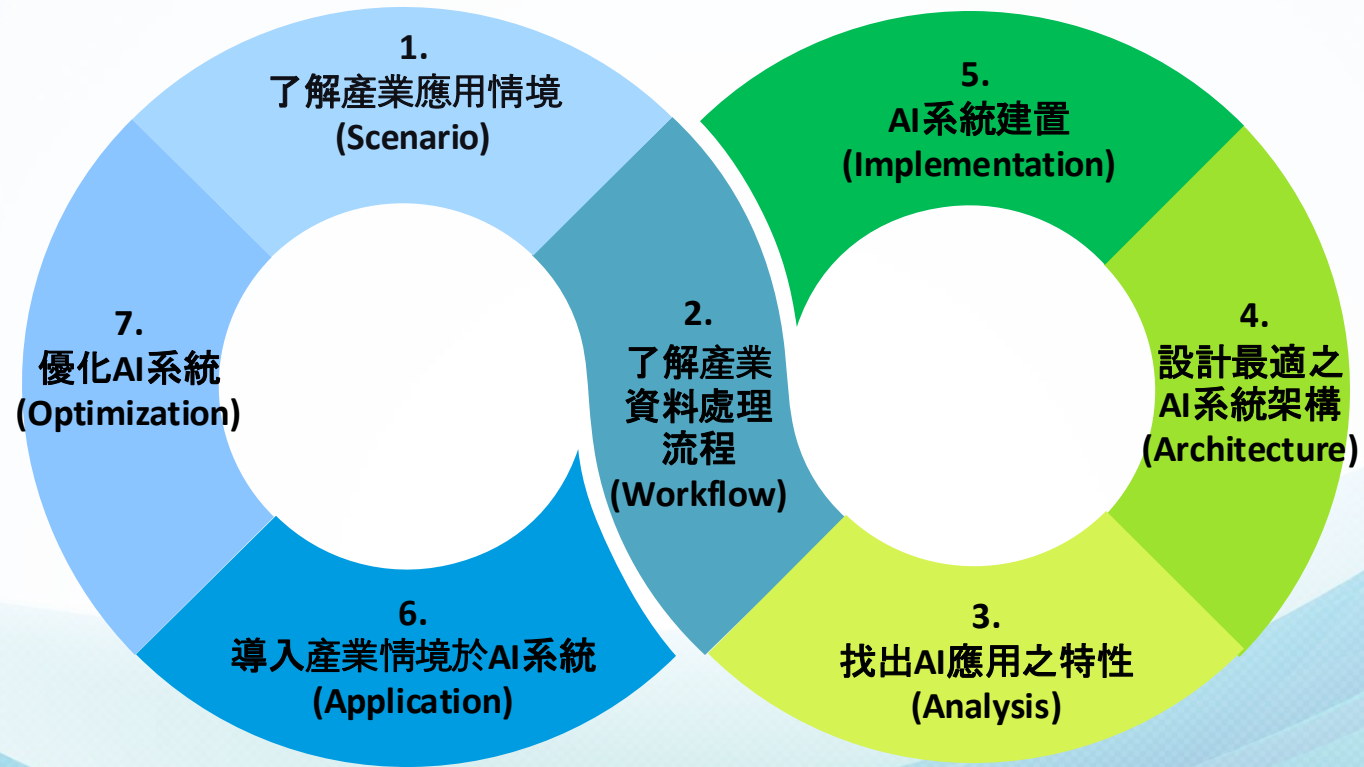
QCT DevCloud 乃是針對高效能運算、資料分析與機器學習等應用服務工作負載，並基於**QCT POD架構**，整合了QCT新世代運算平台、預編譯HPC/AI應用程式及完整軟體開發工具環境之**開放式平台**，提供使用者藉由遠端快速存取以進行相關應用程式開發評估、整合測試與優化



平台特點

- ✓ 針對HPC/AI/DA應用的混合式運算平台
- ✓ 支援Bare Metal及Cloud-native作業環境，提供自動化資源與工作排程管理
- ✓ 提供叢集系統管理及即時系統監控工具
- ✓ 採用軟體定義儲存系統及階層式儲存管理機制

與時俱進的現代化AI系統



以全面性的思維，找出產業應用之特性，結合最適AI關鍵要素
並不斷進行優化基礎架構，以引領產業創新



www.qct.io

 **Thank You** 

