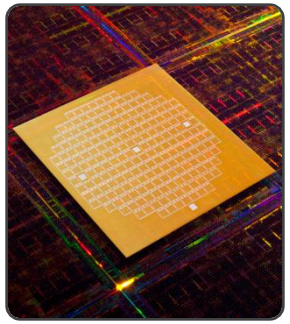
NCHC End2End LLM Bootcamp & Taiwan AI RAP Platform

June 17-18, 2025

NCHC X OpenACC X NVIDIA

CUDA-X Accelerates Every Industry

6M+ Developers & 900+ SDKs/Models

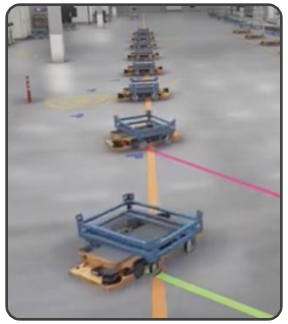


cuLitho

Computational Lithography

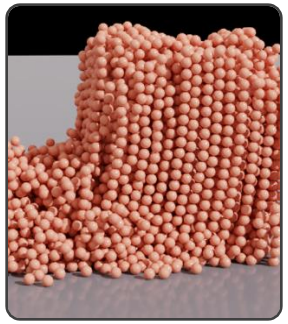
cuDSS

CAE



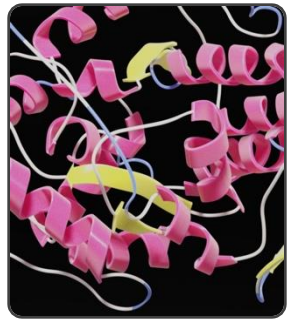
cuOpt

Decision Optimization



Warp

Physical Simulation



cuDF

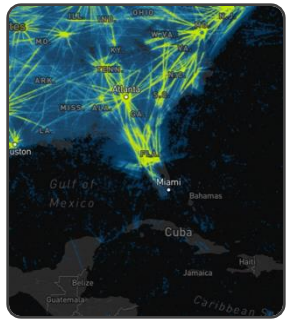
Data Processing

PhysicsNeMo

AI Physics

CUDA-Q

Quantum Computing

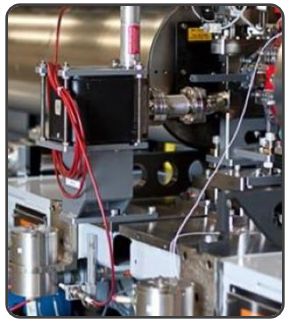


cuEquivariance

Drug & Materials Discovery

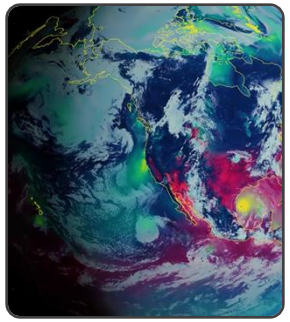
ALCHEMI

AI Materials Science



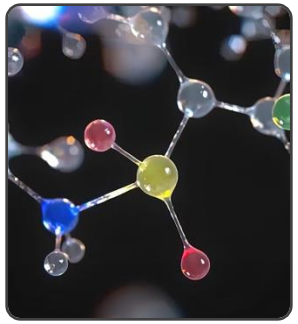
Holoscan

Edge HPC



Earth-2

Weather Analytics



Parabricks

Gene Sequencing

cuPyNumeric

Numerical Computing

NCHC-NVIDIA Joint Lab

https://github.com/nqobu/nvidia/

Repository for NCHC-NVIDIA Joint Lab

- tutorials - NVIDIA online courses/tutorials in AI/HPC
- 20210412 - NVIDIA Techniques Sharing 2021
- 20210706 - NCHC Techniques Sharing 2021
- 20211202 - NVIDIA Techniques Update 2021
- 20211221 - AI+HPC: 利用 NVIDIA Modulus 實踐 PINN 於物理模擬
- 20220415 - NVIDIA Techniques Update 2022
- 20220530 - NCHC-NVIDIA Techniques Sharing 2022
- 20220629 - PINN 與 NVIDIA Modulus 實作訓練營
- 20221111 - Quantum Computing Workshop / 量子計算模擬實作
- 20230413 - NVIDIA Techniques Sharing 2023
- 20230517 - NVIDIA Techniques Briefing: NVIDIA Federated Learning
- 20230525 - AI for Science: NVIDIA Modulus 及 NVIDIA Omniverse 實作
- 20230727 - N-Way to GPU Programming Bootcamp / 多 GPU 程式設計訓練課程
- 20230821 - NVIDIA Techniques Salon 2023: Programming the NVIDIA Superchip
- 20231207 - NCHC Open Hackathon 2023
- 20240410 - NCHC Quantum Computing Bootcamp 2024 - NVIDIA CUDA-Q and cuQuantum
- 20240506 - AI for Science: NVIDIA Modulus, NVIDIA Omniverse, and NVIDIA Earth-2
- 20240508 - NCHC Techniques Sharing 2024
- 20240626 - NCHC AI for Science Bootcamp 2024 - NVIDIA Modulus 物理模擬計算
- 20240806 - NCHC End-to-end LLM Bootcamp 2024 - NVIDIA NeMo 大型語言模型框架
- 20240924 - NCHC N-Way Bootcamp 2024 – NVIDIA GPU 加速運算
- 20241114 - Earth-2 Overview
- 20241129 - NCHC × NTU - NVIDIA BioNeMo Protein Design Workshop 2024
- 20241204 - NCHC Open Hackathons 2024
- 20250218 - NCHC Grace Workshop 2025
- 20250415 - NCHC N-Way Bootcamp 2025 – NVIDIA GPU 加速運算

Co-innovate with Developers

3 Ways

CUDA-X Bootcamp	Open Hackathon	NVAITC Projects
Training	Acceleration	Collaboration

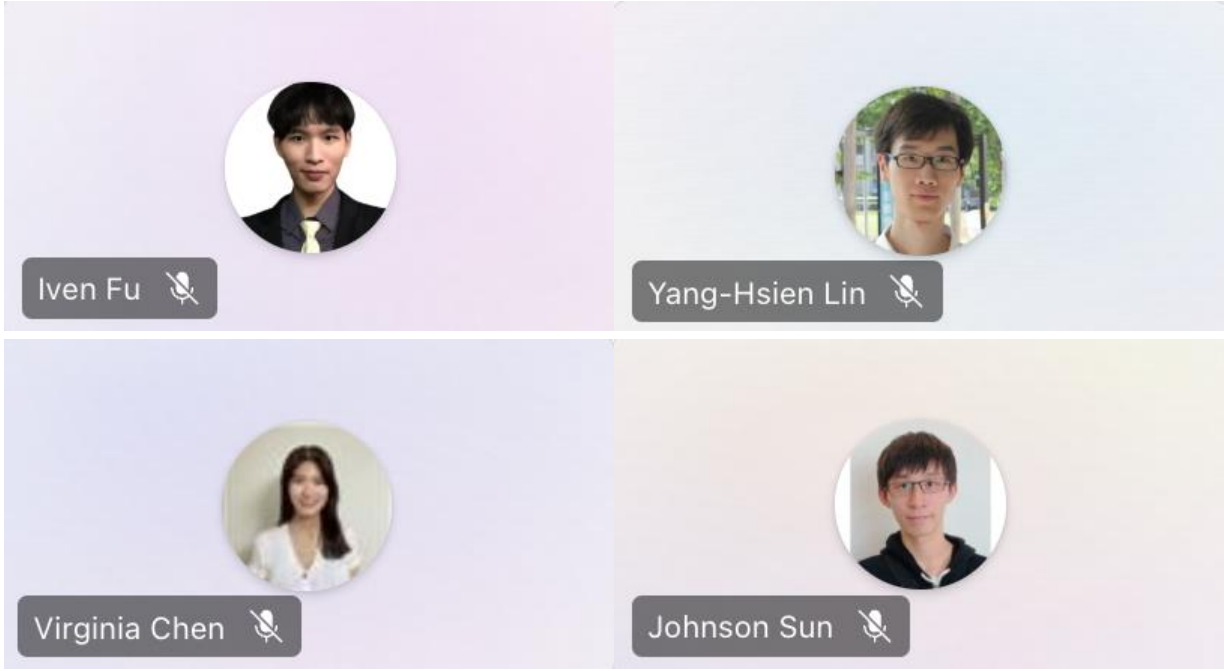
Co-innovate with Developers

3 Ways

CUDA-X Bootcamp	Open Hackathon	NVAITC Projects
Training	Acceleration	Collaboration

NVIDIA NeMo Bootcamp - Team Roster

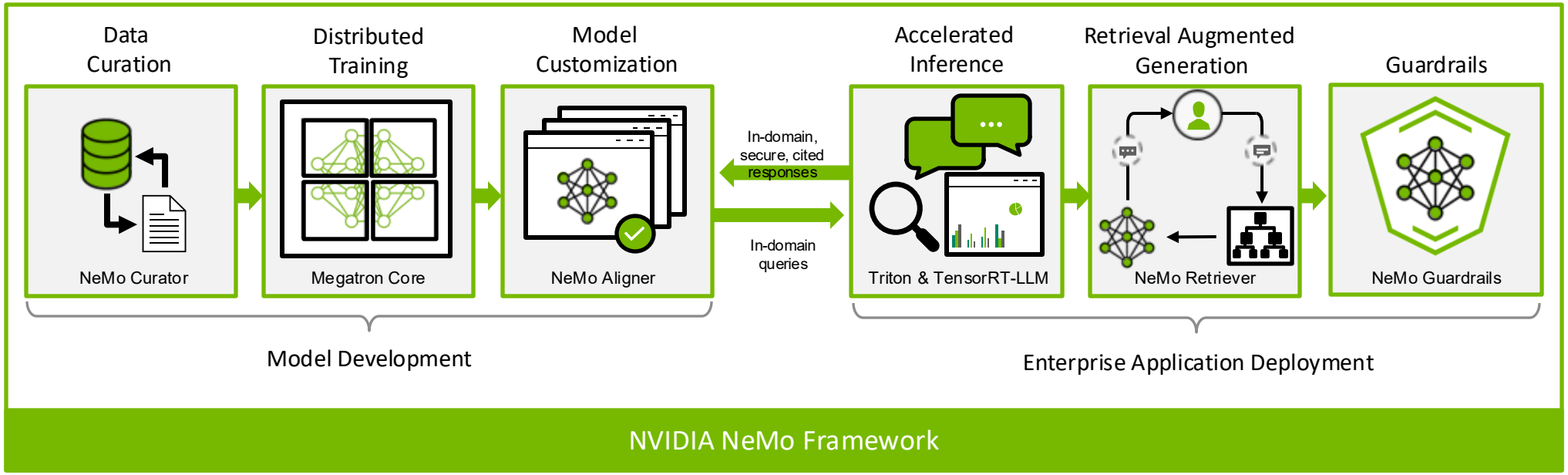
Instructors



Teaching Assistants

Building an End-to-End Generative AI

Build, customize and deploy generative AI models with NVIDIA NeMo



NCHC Taiwan AI RAP Platform & Sovereign AI @ GTC Taipei 2025

https://www.nvidia.com/en-us/on-demand/session/gtctpe25-stw51018/

在全球 AI 競賽中，主權 AI 已成為國家科技發展的重要戰略。值得信賴的 AI 對話引擎 (Trustworthy AI Dialogue Engine，TAIDE) 致力於發展以台灣本土價值為基礎的語言模型，以保護台灣文化與價值觀、推動符合在地需求的技術與應用。本演講將探討主權 AI 的核心價值、TAIDE 的發展願景、應用現況、當前面臨的挑戰，以及政府、企業與學術界如何攜手推動主權 AI 的應用。

LLM Bootcamp - NVIDIA NeMo 大型語言模型訓練實戰教學

https://github.com/wcks13589/LLM-Tutorial

LLM Bootcamp - NVIDIA NeMo 大型語言模型訓練實戰教學 🚀

歡迎來到 LLM Bootcamp！本教學將帶您完整體驗使用 NVIDIA NeMo 進行大型語言模型（LLM）的完整流程，從零開始學會模型轉換、預訓練、微調到部署的實戰技巧。

🎯 學習目標

通過本 Bootcamp，您將學會：

1. 🔄 **模型轉換技能**：掌握 Hugging Face 與 NeMo 格式間的轉換
2. 📚 **預訓練實踐**：體驗大規模語言模型的持續預訓練
3. 🛠️ **微調技術**：學會針對特定任務進行模型微調、掌握 LoRA 等參數高效微調方法
4. 📊 **模型評估**：學會評估和測試模型性能
5. 🚀 **模型部署**：了解模型導出和部署流程

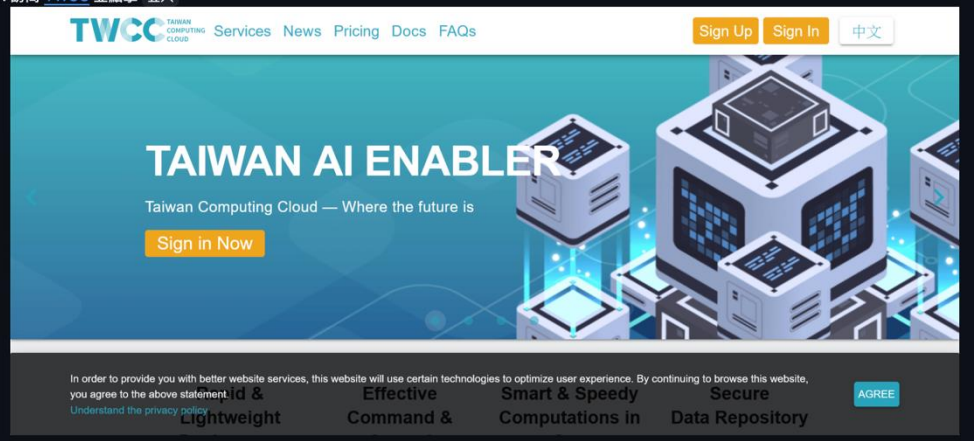
📁 教學大綱

- 🚀 開始之前：環境設定
  - 📖 詳細環境設定指南 ⭐
- 📩 專案設置
- 📖 詳細教學步驟
  - 第一章：模型轉換基礎
  - 第二章：持續預訓練實戰
  - 第三章：指令微調技術
  - 第四章：Reasoning 資料微調技術
  - 第五章：模型評估與測試
  - 第六章：模型部署與轉換
- 💡 實戰技巧
- 📚 進階學習資源

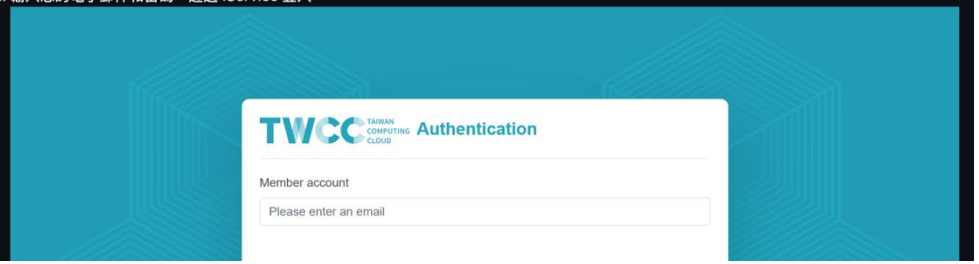
📋 建立 TWCC 容器

步驟 1：登入 TWCC

1. 訪問 TWCC 並點擊 登入



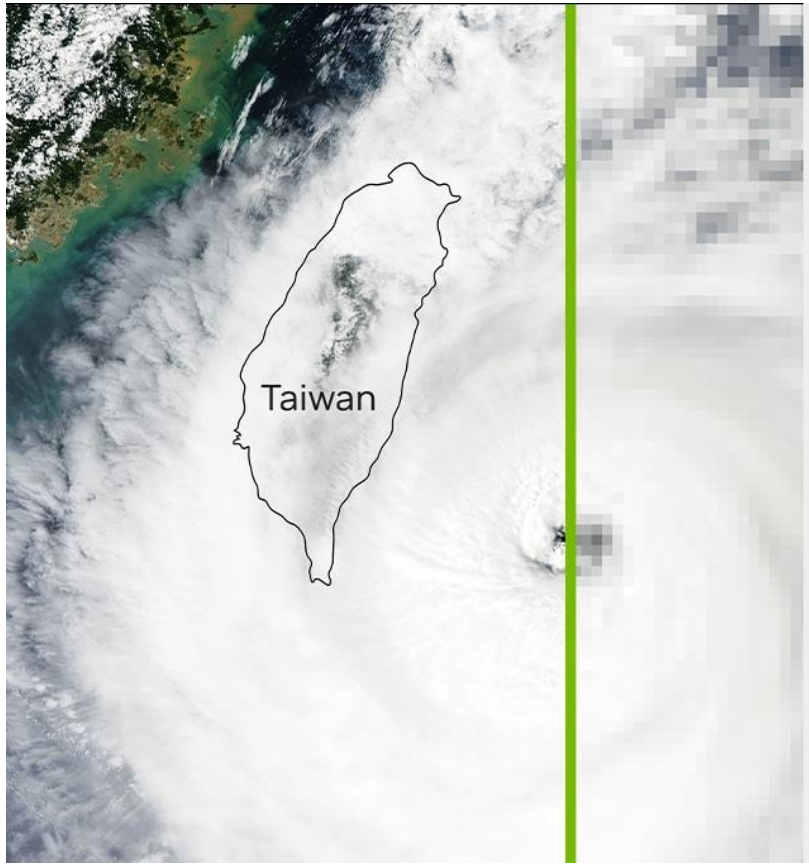
2. 輸入您的電子郵件和密碼，透過 iService 登入



NVIDIA 驅動的超級電腦為台灣研究帶來大躍進

1700+ H200 + 2 GB200-NVL72

https://blogs.nvidia.com.tw/blog/taiwan-research-supercomputer/



NVIDIA GB200 NVL72

Delivers New Unit of Compute

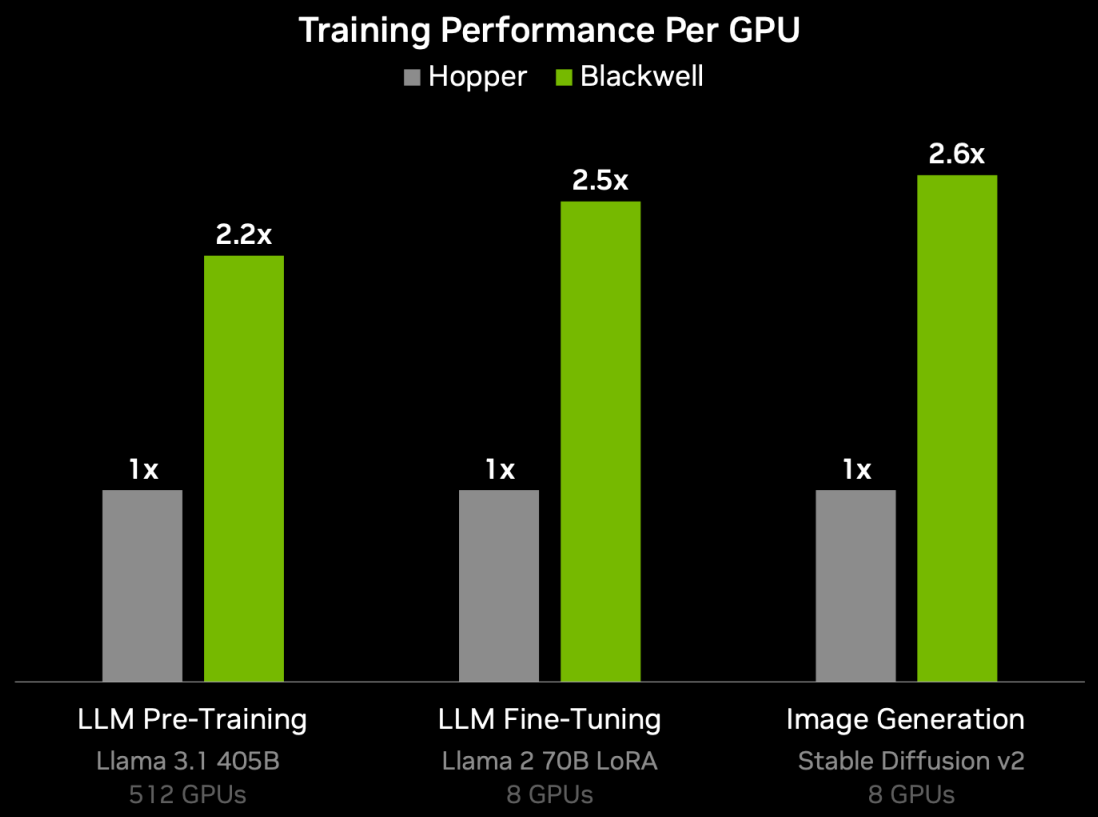
GB200 NVL72

- 36 GRACE CPUs
- 72 BLACKWELL GPUs
- Fully Connected NVLink Switch Rack

HPC FP64	2.88 PFLOPs
Training FP8	720 PFLOPs
Inference FP4	1,440 PFLOPs
NVL Model Size	27T params
Multi-Node All-to-All	130 TB/s
Multi-Node All-Reduce	260 TB/s

Blackwell Over 2.5X Hopper Training Performance

First available-category submissions using GB200 NVL72 rack-scale architecture

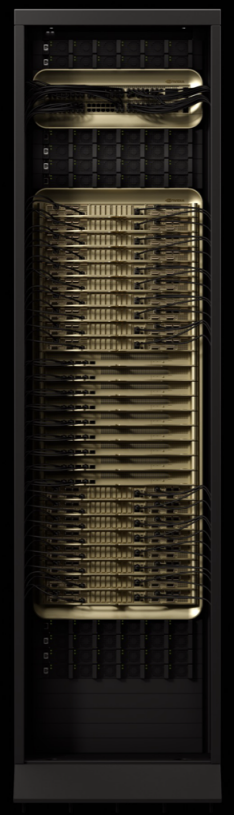


MLPerf Training v5.0 Closed. Results retried on June 4, 2025, from www.mlcommons.org, from the following entries: 5.0-0014, 5.0-0071, and 5.0-0076 . Hopper results using H100 GPU for LLM Fine-Tuning and Image Generation using results from MLPerf Training v4.1 from result ID 4.1-0050. Results verified by MLCommons Association. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

NVIDIA

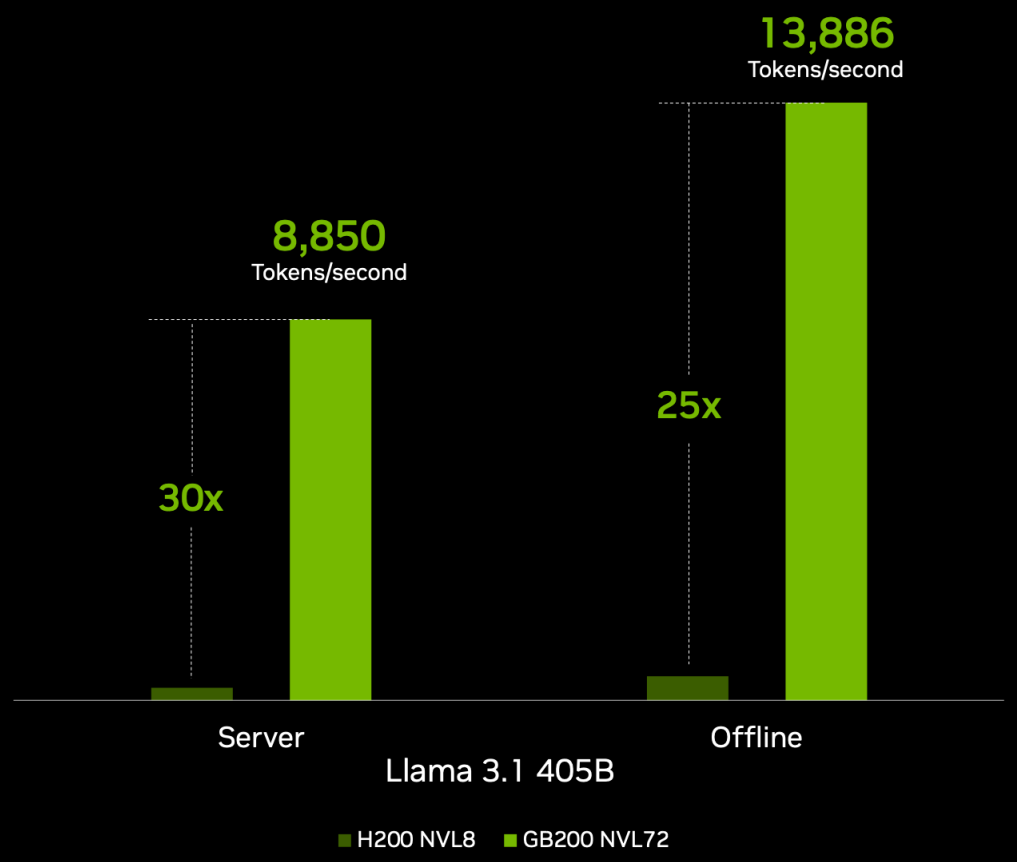
GB200 NVL72 Increases Token Throughput by 30x

New Llama 3.1 405B benchmark record



GB200 NVL72

72 Blackwell GPUs
36 Grace CPUs
13.4 TB HBM3e | 576 GB/s
130 TB/s NVLink



MLPerf Inference v5.0, Closed, Data Center. Results retrieved from www.mlcommons.org on April 2, 2025. Results retrieved from the following entries: 5.0-5.0-0058, 5.0-0060. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorize use strictly prohibited. See www.mlcommons.org for more information

NVIDIA

Co-innovate with Developers

3 Ways

CUDA-X Bootcamp	Open Hackathon	NVAITC Projects
Training	Acceleration	Collaboration

OpenACC Open Hackathon

Celebrating 12 years and continue building the communities.





Ecosystem Development

Training/Education

OpenACC Specification

TechOrange 科技橘

Podcast | Youtube | 聯絡我們 | 活力徵才 搜尋...

第五屆智動化大學 AI 創新顛覆傳統製造 必修 TOP3

最新文章 趨勢專題 全新一週 TO Highlight 主題分類 AI 人工智慧 智慧製造 資安 產業 AI 治理/公共 人才/職場 產業分析

AI 人工智慧

在 3 週內實現高達萬倍的運算效能提升！NCHC、NVIDIA、OpenACC「NCHC Open Hackathon」黑客松，提供開發者實現 AI 創新最佳平台

NVIDIA 2025-03-13

https://buzzorange.com/techorange/2025/03/13/nchc-open-hackathon-nvidia/

OpenACC More Science, Less Programming | OPEN HACKATHONS | NCHC | NIAR 國家實驗研究院 國家高速網路與計算中心 National Center for High-performance Computing | NVIDIA

2024/11/13-12/04: Open Hackathon (12 teams)

Team	Mentor	Core Area of Focus	Domain	Languages/ Libs	How much Speedup?	Why acceleration matters?
1-Dream Chaser	Anthony Chang (Engrg-Hardware 1) TW Ying-Ja Chen TW	Protein-small molecule docking	Healthcare Bioinformatics	CUDA	7.8X	Compared to the original AutoDocker-GPU, OmegaDocker supports larger molecular docking simulations. In addition to its high computational intensity, it also demands greater bandwidth.
2-NYCU HPC team2	Shijie Wang CN	Accelerate NVLM 1.0 inference	LLM Multimodal	python/pytorch	41.7X	Multimodal large language models face numerous challenges in inference acceleration, including high computational resource consumption and slow response times. By leveraging the latest inference acceleration technologies, GPU computing power can be fully utilized, effectively reducing inference latency, improving interactive experiences, and expanding application scenarios.
3-氣象署-興大應數聯隊	Leo Chen (Engrg-Hardware 1) TW	Accelerate physics parameterization in weather forcasting model	Weather	Fortran	70.6X	The global weather forecasting model TCo is divided into the dynamical core (GPU) and physical parameterization (CPU). Computationally intensive and data-independent subroutines are ported to the GPU to effectively utilize its computing power.
4-NTUT_BirdSong	Virginia Chen TW Iven Fu TW	Accelerate audio foundation model pretraining	Audio	python/transformer_engine	3.6X	To create a foundation model for bird songs, pre-trainning speed really matters.
5-Parallel Minds	Reese Wang TW	Accelerate firefly algorithm	HPC	CUDA	9X	The Firefly Algorithm has a wide range of applications. It can be applied not only to path prediction (navigation) but also to optimizing renewable energy systems (maximizing solar cell efficiency), gene regulatory network modeling, drug design, image processing, and more.
6-NTHU_LSALAB	Kevin Chen TW Sungta Tsai TW	Acclerate inference process performance	DPU	C/DOCA	1.23X	In current inference and model computations, the CPU is responsible for controlling and transmitting the data to be processed, which limits the GPU's computational power due to data handling speeds. By utilizing DPU I/O to directly access GPU memory, the processing capabilities of inference and models are enhanced.
7-NoLab	Pika Wang TW Ikko Hamamura JP Tian Zheng (Engrg-Hardware 1) CN	Variational quantum eigensolver (VQE)	Quantum Chemistry	CUDA-Q	8282X	The Variational Quantum Eigensolver (VQE) is a promising quantum algorithm for determining the optimal ground-state energy of molecules, a fundamental of chemical reactions and drug discovery. Efficient VQE simulation can help design quantum algorithms, accelerate scientific research, and reducing development cycles.
8-Elsa Robotics	Johnson Sun TW Frank A. Lin (Engrg-Hardware 1) TW	Accelerate robotics navigation pipeline	Robotics	Python/CuPy/TensorRT	11X	Low latency and high frame rates are critical for a robot's real-time responsiveness and operational precision in dynamic environments. Relying solely on CPU computation makes it difficult to meet real-time processing demands; therefore, GPU assistance is required in specific computational stages to enhance frame rates.
9-GBA-VVM	Min Yu CN Leo Chen (Engrg-Hardware 1) TW	Accelerate advection subroutine	Weather	Fortran/OpenACC	18X	The current atmospheric forecasting accuracy is typically at the kilometer scale, while VVM has improved it to the meter scale. This requires extensive computational resources.
10-smile lab	Ken Liao Yang-Hsien Lin TW	Federated Learning for Pathology	Healthcare Histopathology	Python/cuCIM and Pytorch Lightning	5.5X	Whole Slide Images (WSIs) are massive, making patch extraction computationally intensive. Efficient extraction is key for preprocessing in federated learning, where distributed nodes handle large datasets. Accelerating this process with optimized pipelines (e.g., caching, GPU acceleration) reduces preparation time and avoids I/O-related performance losses.
11-Plantmen	Cliff Chiu TW	Acclerate RAG inference pipeline performance	LLM Multimodal RAG	Python, TensorRT	10X	The extisting pipeline for RAG is too slow for the user to get the response on the LINE chatbot.
12-CYCU_Quantum	Pika Wang TW Ikko Hamamura JP Anderson Meng TW	Quantum PageRank	Quantum Machine Learning	CuPy	10000X	Quantum PageRank leverages quantum interference to reveal complex relationships between nodes, enhancing rankings in applications for websites and social networks. However, simulating noisy interference is much more complex than typical statevector simulation (2^2n vs. 2^n). Further speedup is required to investigate Quantum PageRank in real-world scenarios, which has not yet been achieved.

Other Publications

- 工商: https://www.ctee.com.tw/news/20250408700939-431204?utm=LINE_share_btn
- 經濟: https://money.udn.com/money/story/5635/8659767
- 引新聞: https://innews.com.tw/223620/
- Line: https://today.line.me/tw/v2/article/yzvjnGz

NVIDIA Developer Program

NVIDIA 開發者計畫

Program Benefits:

Tools

- 900+ exclusive SDKs and models
- GPU-optimized software, model scripts, and containerized apps
- Early access programs

Training

- Research papers, technical documentation, webinars, blogs, and news
- Technical training and certification opportunities
- 1,000s of technical sessions from industry events On-Demand

Community

- NVIDIA developer forums
- Exclusive meetups, hackathons, and events

Special Program (Present to Jan 2024)

- Join NVIDIA Developer program now, you will get one NVIDIA Training

Join the Community

Reference Links

- [NVIDIA NeMo LLM Bootcamp Tutorials]
  - https://github.com/wcks13589/LLM-Tutorial
  - https://github.com/NVIDIA/NeMo
- [NCHC Taiwan AI RAP]
  - https://www.nvidia.com/en-us/on-demand/session/gtctpe25-stw51018/
  - https://rap.genai.nchc.org.tw/
- [NCHC New Supercomputer]
  - https://blogs.nvidia.com.tw/blog/taiwan-research-supercomputer/
- [MLCommons MLPerf Benchmark Results]
  - https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/
- [NVIDIA Developer Program]
  - https://developer.nvidia.com/developer-program
- [NCHC Open Hackathon 2024]
  - https://www.nchc.org.tw/Message/MessageView/3949?mid=46&page=1
- [NCHC NVIDIA Joint Labs]
  - https://github.com/nqobu/nvidia/
- [Today's Bootcamp Feedback Survey]
  - https://forms.office.com/r/zCRcC3Av8D

Feedback Survey
https://forms.office.com/r/zCRcC3Av8D

June 17-18 NCHC End-to-End LLM Bootcamp Feedback Survey

Thank You