2024 NCHC Open Hackathon

NCHC X OpenACC X NVIDIA

OpenACC
More Science, Less Programming

Agenda

2024 / 11 / 13 (Wed) ： Day 0: Kick-off Meeting 14:00~16:00 PM (Online)

- 02:00 PM - 02:05 PM: Welcome & Opening
- 02:00 PM - 02:15 PM: Event Overview
- **02:15 PM - 03:15 PM: Team self-introduction and getting to know the mentor.**
  - 1 min for all mentors per team
  - 3 mins for each team lead
- 03:15 PM - 03:30 PM: Introduction to computing resources
- 03:30 PM - 04:00 PM: Team discussion
  - **Before** the event, Please self-study the profiling training materials, and see more here.
  - **During** the event, Team discussion on profiling techniques. (Each team's mentors will create a meeting link)
  - **After** the event, Post general profiling questions in #profiler-support.

Team Roster

Group Green (Hosted by Jay)

Group Blue (Hosted by CK)

- Host: Jay Chen & Dr. CK Lee
- Infra: Johnson Sun
- NCHC Contact Window: Zhoujin Wu
- Ack NCHC, OpenACC and NVIDIA Team
  - Vincent Chiu
  - Ryan Jeng
  - Bharat Kumar
  - Apoorva Laharia
  - Yash Gupta
  - Elica Kyoseva
  - Eddie Huang
  - Eric Kang
  - Kuan-Ting Yeh
  - Jinny Lin

Team-ID	Team	Mentor	Domain	Host
1	Dream Chaser	**Anthony Ying-Ja**	Healthcare Bioinformatics	Jay
2	NYCU HPC team2	**Shijie**	LLM Multimodal	Jay
3	氣象署-興大應數聯隊	**Leo**	Weather	Jay
4	NTUT_BirdSong	**Virginia Iven**	Audio	Jay
5	Parallel Minds	**Reese**	HPC	CK
6	NTHU_LSALAB	**Kevin SungTa**	DPU	CK
7	NoLab	**Pika Ikko Tian**	Quantum Chemistry	CK
8	Elsa Robotics	**Johnson Frank**	Robotics	CK
9	GBA-VVM	**Min**	Weather	Jay
10	smile lab	**Ken Yang-Hsien**	Healthcare Histopathology	CK
11	Plantmen	**Cliff**	LLM Multimodal RAG	Jay
12	Quantum Walk	**Pika Ikko Anderson**	Quantum Photosynthetic	CK

OPENACC – CELEBRATING 11 YEARS

Building Community.





Ecosystem Development

Training/Education

OpenACC Specification

OAK RIDGE National Laboratory · AMD · NVIDIA · arm · NCAR · SIEMENS Ingenuity for life · HZDR Helmholtz Zentrum Dresden Rossendorf · ETH zürich · University of Delaware · Indiana University · Berkeley Lab · NOAA

大阪大学 Osaka University · 東京工業大学 Tokyo Institute of Technology · Stony Brook University · epcc · Illinois · Total · CDAC Centre for Development of Advanced Computing · Brookhaven National Laboratory · SUSE · 筑波大学 University of Tsukuba

Michigan State University · RIKEN R-CCS · University of Houston · VirginiaTech · KAUST · NCHC · Technische Universität Dresden

Open Hackathon Objectives

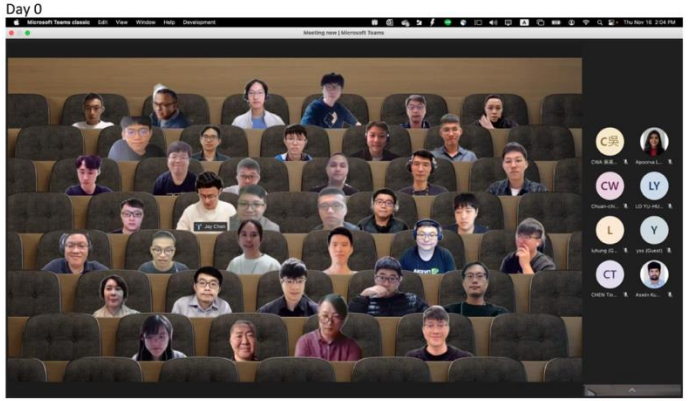
Connect	Accelerate	Celebrate
Developers & Mentors Apps & Acceleration	Speedup Energy Efficient	Publication, Co-Paper Blogs and Talks

Please ack NCHC and Mentors if any publications.

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

2023 NCHC Open Hackathon

Day 0



Final Day – AM Sessions

Highlights: (*Note most benchmarks results come with specific conditions.)

- "468X vs CPU" – Team 2 haofan2023
- "3X by TensorRT-LLM on TAIDE model" – Team 12 NCHC-Speedrunning
- "126X, the CPU implementation is significantly less suitable in time critical applications." – Team 1 Schrödinger s cat
- "Speed up about 8x with only 25 lines modification. CPU cost 10+ months, we never think about re-generate mesh, now we can!" – Team 5 CWA mesh generation
- See below **Team Results/Outcomes** for more!

*Note most benchmarks results come with specific conditions.

Team Name	Mentors	Code Area of Focus	Domain	Languages /Libs	Results	Energy Efficiency	Comments
1 Schrödinger s cat	Reese Yun-Yuan	Quantum-Inspired Algorithm (QUBO)	HPC Optimization	C++ /JAX	126X	126X	The CPU implementation is significantly less suitable in applications where timeliness is critical, such as high-frequency trading or real-time route planning, which demand rapid decision-making within a limited timeframe.
2 haofan2023	Tian Frank Yun-Yuan	Quantum Circuit Simulation (QFT, QAOA)	HPC Quantum	C++	468X (vs 1 core)	65X	Benchmark up to 30 qubits. They could increase up to 39 qubits with special SSD cache optimization innovation.
3 NTHU-LSALAB	Sungta Erez	5G SBA (Service Based Architecture)	HPC Network Infra	C++ /DOCA	DPU Save CPU 15% reduce 11X latency	N/A	Utilizing the natively provided high-efficiency SR-IOV network architecture to increase the network speed of 5G Service. Preserving a significant amount of CPU processing resources for 5G SBA components.
4 NTUST CFD Lab	Bharat Shijie Kuan-Ting	3D-CFD (Direct Forcing Immersed Boundary, LES turbulence model)	HPC CFD	C++ /OpenACC	16.7X	46X	This marks another milestone for our in-house code as large-scale computations are required to simulate real-world cases. The significant speedup achieved is a substantial leap forward for us in reaching that goal.
5 CWA mesh generation	Leo Jay	Mesh generation for MPAS model	HPC CWO	Fortran /OpenACC	12X (vs pre-event)	22X	Speed up about 8x with only 25 lines modification. High-res mesh generation takes estimiated 10+ months, now <1 month with GPU, which is incredible. we never think about re-generate mesh, now we can!
6 CYCU BME	Eason	Otoscopy Diagnosis	AI Healthcare	/Pytorch TensorRT	1.5X	N/A	Model inference optimization implementation based on tensorRT.
7 CWA_GVER	Ming Kuan-Ting Jay	Global Ensemble model Verification	HPC CWO	python /JAX	44X (wo I/O)	N/A	Verification costs 12+ hours, which is not feasible with CPU only in future operation, not to mention higher resolution and more ensemble members.
8 WTMH	Ying-Kai	Arrhythmia Screening of Real-Time Single-Lead ECG	AI Healthcare	/Pytorch TensorRT Triton	40X	N/A	Achieved "instant" ECG analysis system by TensorRT+Triton.
9 氣興聯隊	Leo Jay	CWAGFS-TCo	HPC CWO	Fortran /OpenACC	1.8X (wo I/O)	1.6X	The first and only in-development gpu-accelerated NWP in CWA.
10 YSS	CK Tian Frank	Functional Encryption (BSGS) apply in Machine Learning Service	AI Data Privacy	C++	2.2X	3.1X	It is important for server to provide fast and secure service, Functional Encryption is one of the cryptosystem mechanisms used for data privacy In MaaS.
11 TXM-AI	Warren	X-ray Background Correction Model	AI Healthcare	/MONAI	30X	21X	This will save considerable acquisition time and enhance the efficiency of nano CT scans, ultimately accelerating progress in paleontology, biomedical research, materials science, energy, and other related fields.
12 NCHC-Speedrunning	Anthony Cliff	LLM inference with TensorRT-LLM on NCHC servers	AI LLM	/TensorRT-LLM	3X (bs=1)	N/A	Test on TAIDE Model, Taiwan ChatGPT project, based on LLaMA2-7B. could be further improved on H100 in near future.

NARLabs 國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing

(extend from 2022 Open Hackathon)

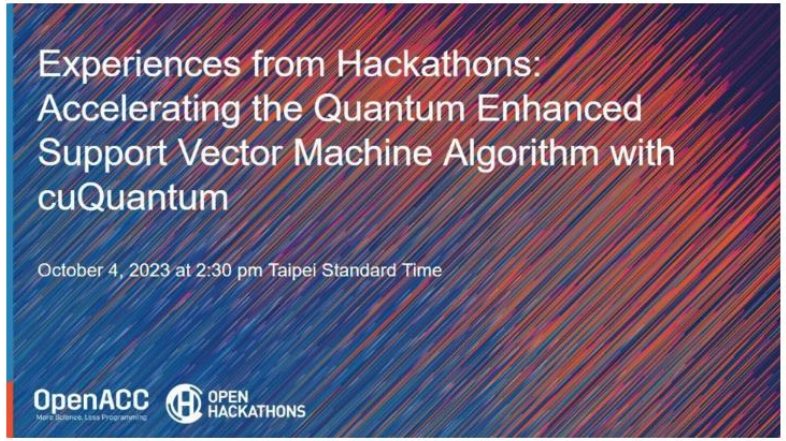
cuTN-QSVM

https://arxiv.org/abs/2405.02630, 4th May 2024.
https://github.com/Tim-Li/cuTN-QSVM

國家同步輻射研究中心
National Synchrotron Radiation Research Center
Imperial College London

Wednesday Oct 4, 2023
02:30 - 02:50 PM

Experiences from Hackathons: Accelerating the Quantum Enhanced Support Vector Machine Algorithm with cuQuantum



Speaker

Tai Yue Li
National Synchrotron Radiation Research Center (NSRRC)

cuTN-QSVM: cuTensorNet-accelerated Quantum Support Vector Machine with cuQuantum SDK

Kuan-Cheng Chen**
QuEST
Imperial College London
London, United Kingdom
kc2816@ic.ac.uk

Tai-Yue Li**
National Synchrotron Radiation Research Center
Hsinchu, Taiwan
li.timty@nsrrc.org.tw

Yun-Yuan Wang**
NVIDIA AI Technology Center
NVIDIA Corp.
Taipei, Taiwan
yunyuanw@nvidia.com

Simon See
NVIDIA AI Technology Center
NVIDIA Corp.
Santa Clara, CA, USA
ssee@nvidia.com

Chun-Chieh Wang
National Synchrotron Radiation Research Center
Hsinchu, Taiwan
wang.jay@nsrrc.org.tw

Robert Willie
Chair of Design Automation
Technical University of Munich
Munich, Germany
robert.wille@tum.de

Nan-Yow Chen
National Center for HPC
Narlabs
Hsinchu, Taiwan
nanyow@nchc.narl.org.tw

An-Cheng Yang
National Center for HPC
Narlabs
Hsinchu, Taiwan
1203087@nchc.narl.org.tw

Chun-Yu Lin
National Center for HPC
Narlabs
Hsinchu, Taiwan
lincy@nchc.narl.org.tw

README | Code of conduct | Apache-2.0 license

License Apache 2.0 | arXiv 2405.02630

cuTN-QSVM: cuTensorNet-accelerated Quantum Support Vector Machine with cuQuantum SDK

Welcome to the official repository of cuTN-QSVM, featuring fast GPU simulators for benchmarking Quantum Support Vector Machines (QSVMs) and scripts for generating compatible quantum circuits for hardware execution. Facilitated by NVIDIA's cuQuantum SDK and the cuTensorNet library, this project integrates cutting-edge quantum computing technologies with high-performance computing systems, enhancing quantum machine learning's efficiency and scalability to new heights.

Project Overview

Quantum Support Vector Machines (QSVMs) utilize a quantum-enhanced approach to tackle complex, multidimensional classification problems, surpassing the capabilities of classical SVMs under certain conditions. However, prior to the advent of large-scale quantum systems, the scalability of simulating QSVMs on CPUs was traditionally limited by the exponential growth in computational demands as qubit counts increased. By employing NVIDIA's cuQuantum SDK and the cuTensorNet library, cuTN-QSVM effectively reduces this computational complexity from exponential to quadratic. This enables the simulation of large quantum systems of up to 784 qubits on the NVIDIA A100 GPU within seconds.

Mentored by NVIDIA Pika Wang, Leo Fang, Jay Chen

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

(extend from 2023 Open Hackathon)

Queen: A quick, scalable, and comprehensive quantum circuit simulation for supercomputing

https://arxiv.org/abs/2406.14084

Welcome to compare to our state-of-the-art simulator released in 2024 or offer paid consultation for performance improvement.

Chuan-Chi Wang
National Taiwan University
Taipei, Taiwan
d10922012@ntu.edu.tw

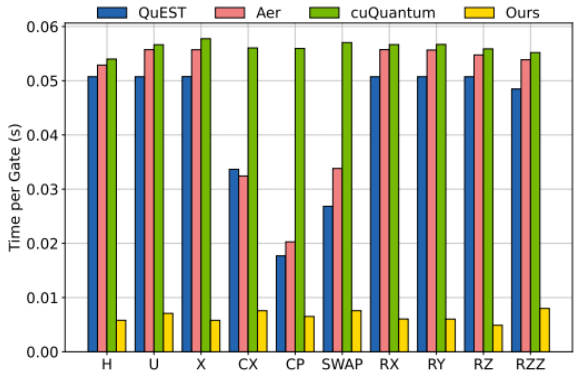
Yu-Cheng Lin
National Taiwan University
Taipei, Taiwan
r11922015@csie.ntu.edu.tw

Yan-Jie Wang
National Taiwan University
Taipei, Taiwan
yanjiewtw@gmail.com

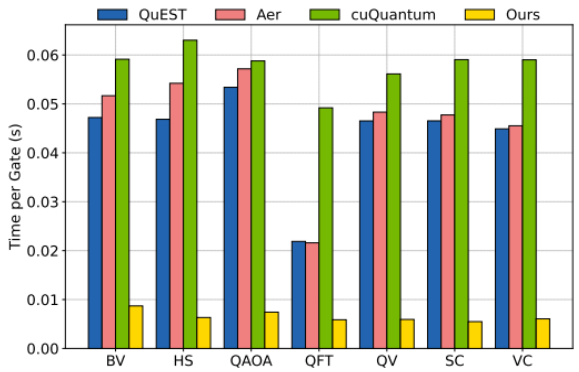
Chia-Heng Tu
National Cheng Kung University
Tainan, Taiwan
chiaheng@ncku.edu.tw

Shih-Hao Hung
National Taiwan University
Taipei, Taiwan
hungsh@csie.ntu.edu.tw

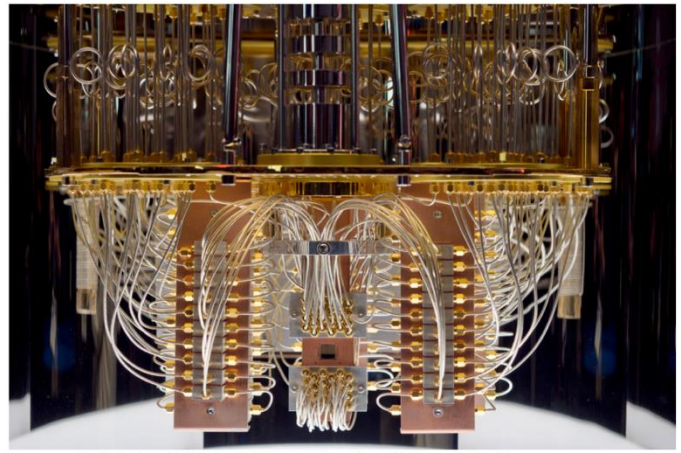
We thank the National Center for High-Performance Computing for providing access to the NVIDIA DGX-A100 workstation.



(a) Gate benchmark.



(b) Circuit benchmark.



haofan2023團隊成員來自臺灣大學資工系「洪士灝老師實驗室」，將量子演算法QAOA加速468倍！

— NVIDIA Mentors: Tian Zheng, Frank Lin, Yun-Yuan Wang

量子技術正以驚人的速度發展，預示著我們即將進入量子計算的時代。在這個過程中，量子電路模擬成為一個關鍵工具，它在量子硬體和軟體的開發中扮演著重要的角色，特別是在處理量子程式的編寫和驗證方面。傳統電腦的強模擬能夠獲得完整的量子狀態信息。這使得傳統電腦在構建量子系統方面變得不可或缺，尤其是在當前嘈雜較多的中等規模量子（NISQ）時代。

量子近似優化算法（QAOA）是一種常用的量子算法，用於通過近似解來解決組合優化問題。然而，在虛擬量子計算機上執行QAOA對於解決需要大規模量子電路模擬的組合優化問題而言，會遇到模擬速度較慢的問題。團隊使用數學優化來壓縮量子操作，並結合有效的位元操作進一步降低計算複雜性，透過GPU加速最高獲取468倍的加速效果！

Table 1: The elapsed time of 5-level QAOA (unit: second, double).

Qubit	CPU_{Single}	$CPU_{Mutiple}$	CPU_{Cache}	GPU_{Cache}	GPU_{All}
23	29.80	1.28 (23x)	1.28 (63x)	0.24 (120x)	0.06 (**341x**)
24	68.00	3.46 (20x)	3.46 (43x)	0.55 (123x)	0.12 (**382x**)
25	152.52	15.32 (10x)	15.31 (45x)	1.19 (127x)	0.23 (**404x**)
26	330.69	33.83 (10x)	33.83 (56x)	2.60 (126x)	0.56 (**417x**)
27	712.26	72.66 (10x)	72.66 (54x)	5.59 (127x)	1.08 (**427x**)
28	1556.87	156.52 (10x)	156.52 (54x)	11.96 (130x)	2.17 (**445x**)
29	3325.55	335.09 (10x)	335.09 (49x)	25.73 (129x)	4.45 (**451x**)
30	7226.46	718.33 (10x)	718.33 (47x)	55.20 (130x)	9.22 (**468x**)

更多資訊請看：https://github.com/nqobu/nvidia/raw/main/20231207/Team02.pdf

Mentored by NVIDIA Frank Lin, Tian Zheng, Pika Wang

OpenACC More Science, Less Programming
OPEN HACKATHONS
NCHC
NARLabs 財團法人國家實驗研究院 國家高速網路與計算中心 National Center for High-Performance Computing
NVIDIA

(extend from 2023 Open Hackathon)

NCAR WRF/MPAS Users Workshop 2024

https://www.mmm.ucar.edu/events/workshops/wrf-mpas
https://www.youtube.com/watch?v=DlnuStktdlg&t=15315s

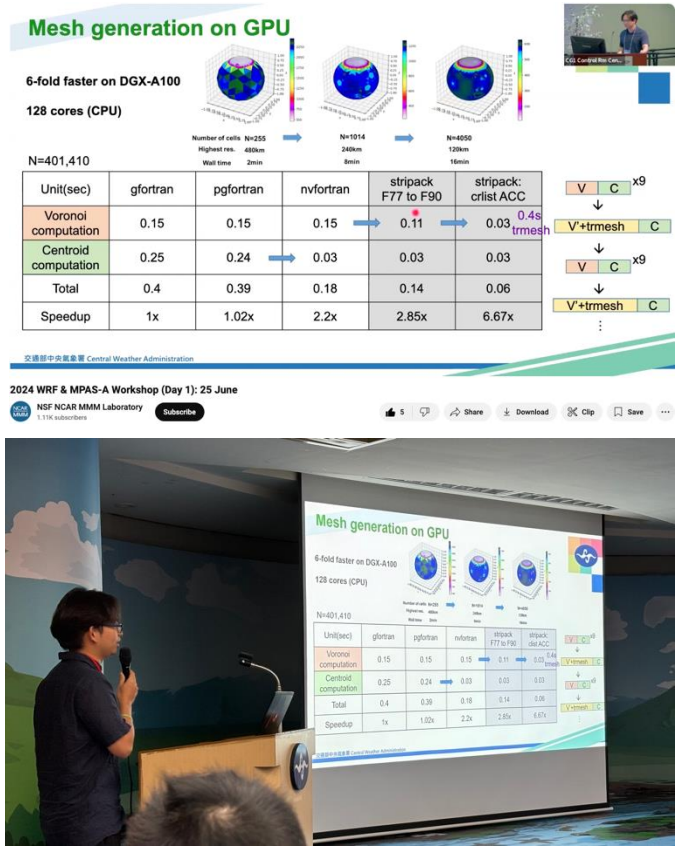
Advancements in Implementing the MPAS-A Regional Model at the Central Weather Administration

[1]Wu, Y.-J., [2]W. Wang, [3]C.-Y. Chen, [3]Y.-L. Chen, [1]S.-L. Huang [1]B.-S. Lin, [1]L.-F., Hsiao
[1]Central Weather Administration, Taiwan
[2]University Corporation for Atmospheric Research, Boulder, Colorado
[3]NVIDIA

In our ongoing efforts to enhance weather modeling capabilities at the Central Weather Administration (CWA), we have integrated several key components from the Weather Research and Forecasting (WRF) operational model into the MPAS-A Regional Model. Collaborating with National Taiwan University (NTU), we have developed the cloud physics scheme TCWA1, which has been integrated into the model framework.

One key focus of our work has been comparing the surface wind speeds simulated by MPAS-A with those from the Weather Research and Forecasting (WRF) model. Our analyses have revealed notable discrepancies between the two models, prompting the introduction of the topo_wind option from WRF. This addition aims to mitigate wind speed biases and improve the overall accuracy of our simulations, particularly in complex terrain regions.

The grid generation program plays a critical role in setting up the computational grid for weather simulations. By leveraging GPU acceleration, we achieved significant improvements in the performance of this program. The tailored GPU acceleration techniques, developed in collaboration with the mentors in NVIDIA workshop, allowed for faster data processing and computation, leading to a six-fold speed increase compared to the previous implementation.



(Present in CWA Conference 2024)

Mentored by NVIDIA Leo Chen, Jay Chen

(extend from 2023 Open Hackathon)

Preserving Data Privacy during Neural Architecture Search for Edge AI Applications

Accepted by ACM RACS 2024

https://www.nchc.org.tw/Message/MessageView/3871?mid=165&page=1

https://github.com/nqobu/nvidia/raw/main/20231207/Team10.pdf

YSS 團隊成員來自台灣大學資工系 洪士灝老師實驗室，將AI加密演算法加速2.2倍！

* NVIDIA Mentors: CK Lee, Tian Zheng, Frank Lin.

新興的AI科技展現了在多個應用領域的潛力。然而，使用使用者收集的資料訓練神經網路模型時，隨之而來的高度隱私敏感性使得隱私問題變得嚴峻。為了保護使用者原始資料，團隊研究Functional Encryption加密技術，但單用CPU無法滿足運算需求，因此團隊將加解密運算移植至GPU加速2.2倍！

CPU Training Time		
Framework	Enc	Training time
LeNet-5	-	4h
FE + LeNet (baseline)	3h	11h33m
parallel enc + parallel training	1h40m	6h6m
Parallel Enc + Opt Dec + Parallel training	1h40m	about 5h

更多資訊請看：

https://github.com/nqobu/nvidia/raw/main/20231207/Team10.pdf

Preserving Data Privacy during Neural Architecture Search for Edge AI Applications*

Yi-Chuan Liang
National Taiwan University
Taipei, Taiwan
d04922003@csie.ntu.edu.tw

Shan-Jung Hou
National Taiwan University
Taipei, Taiwan
r12922146@csie.ntu.edu.tw

Shin-Wei Chiu
National Taiwan University
Taipei, Taiwan
r12922054@csie.ntu.edu.tw

Shih-Hao Hung
National Taiwan University
MBZUAI
Taipei, Taiwan
hungsh@csie.ntu.edu.tw

We acknowledge the fruitful discussions during the 2023 NCHC Open Hackathon with Shao-Fu (Frank) Lin and Tian Zheng from the NVIDIA DevTech Team, and Dr. Cheng-Kuang Lee from the NVIDIA AI Tech Center (NVAITC). We also extend our gratitude to the National Center for High-performance Computing (NCHC) in Taiwan for hosting the 2023 Open Hackathon.

Mentored by NVIDIA Tian Zheng, Frank Lin, Pika Wang.

(extended from 2023 Open Hackathon)

GPU-based Ising Machine for Solving Combinatorial Optimization Problems with Enhanced Parallel Tempering Techniques

https://www.nchc.org.tw/Message/MessageView/3873?mid=165&page=1
https://github.com/nqobu/nvidia/raw/main/20231207/Team01.pdf

GPU-based Ising Machine for Solving Combinatorial Optimization Problems with Enhanced Parallel Tempering Techniques

Kuei-Po Huang*, Chin-Fu Nien*, Yun-Ting Zhang*, Cheng-Kuang Lee†
* Dept. of Computer Science & Information Engineering, Chang Gung University, Taiwan,
† NVIDIA AI Technology Center, NVIDIA Corporation, Santa Clara, CA, USA
Email: * {m1129031, b0929007}@cgu.edu.tw, Corresponding author: watchmannien@mail.cgu.edu.tw, †ckl@nvidia.com

***Abstract*—Ising machines are hardware solvers designed to tackle computationally complex combinatorial optimization problems (COPs), harnessing physical processes such as quantum annealing to simulate the Ising model, enabling these specially designed solvers to tackle a wide range of computationally complex NP-hard problems in real-world applications, such as portfolio optimization, logistics planning, and the traveling salesman problem. While prior works propose to fabricate dedicated integrated circuits for building Ising machines, we leverage off-the-shelf Graphics Processing Unit (GPU) chips for implementing Ising algorithms for quick development. In this work, we explore parallel tempering (PT), an Ising algorithm, which shows promise owing to its capability for concurrent processing of multiple independent searches for the optimal solution, each with a different amount of randomness that allows escaping local minimum. We propose several optimization strategies, including addressing the dependency problem in PT to enhance algorithm parallelism. Our approach incorporates genetic operations that have successfully discovered effective solutions. Empirical evaluations demonstrate that our software solver efficiently finds high-quality solutions for the MAX-CUT combinatorial optimization problem, as evidenced by its strong solution quality on benchmark instances from the well-established G-set**

Tempering (PT) [4], a Markov chain Monte Carlo (MCMC) technique that involves creating multiple replicas of the system, each simulated at a different temperature. By periodically allowing these replicas to exchange their states, PT facilitates the exploration of the system's energy landscape and promotes thermal equilibrium across replicas at different temperatures. This allows the algorithm to escape local minima and explore the energy landscape more thoroughly. However, traditional PT implementations suffer from the sequential nature of the replica exchange process, hindering efficiency and scalability. To address this challenge, we introduce Parallel Quantum-inspired Search (PQS), a novel heuristic designed to efficiently solve QUBO problems across replicas. PQS's algorithm eschews the need for sequential calculation of the entire Hamiltonian by enabling the simultaneous computation of Hamiltonians for all replicas. By coalescing multiple computational steps into a single computational step with parallel issuing of multiple GPU kernels, PQS significantly reduces computational overhead. Furthermore, by leveraging kernel fusion to combine multiple operations into a single GPU kernel, our approach achieves paral-

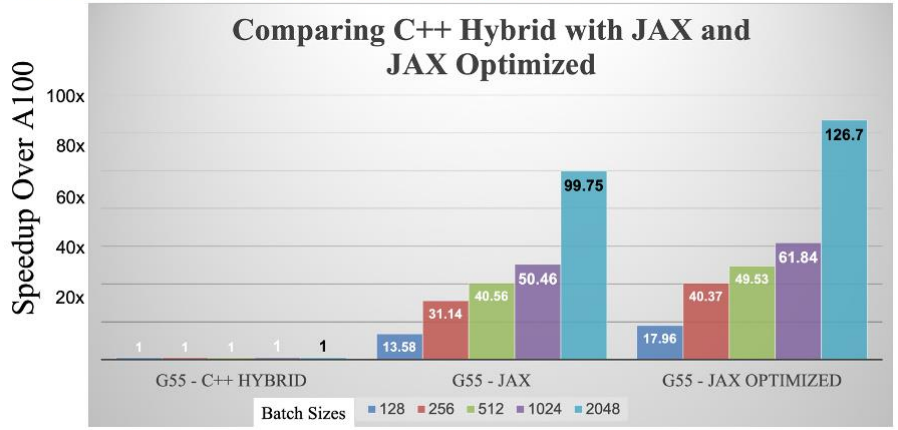
Schrodinger's Cat團隊成員來自長庚大學資工系「粘儆夫老師實驗室」，將量子啟發組合最佳化問題加速126倍！

— NVIDIA Mentors: Reese Wang, Yun-Yuan Wang

在限制條件下找到最佳解，在現實生活中的應用非常廣泛，例如在交通運輸、製造業、金融等領域中都有應用。然而，最佳化問題也很困難，因為它需要處理大量的數據和複雜的限制條件。舉例來說，如果你要在一個城市中設計一個最佳的公交路線系統，你需要考慮到每個站點之間的距離、人流量、交通擁堵情況等多個因素，這就需要用到組和最佳化的方法來找到最佳解。QUBO（Quadratic unconstrained binary optimization）演算法正是解決優化問題的最佳工具之一，可以將現實生活中的問題轉化為數學表示，並通過量子計算等方法快速找到最佳解。這樣，我們就能更快地解決問題，讓生活變得更加便利和高效。

團隊運用了Nsight Systems來識別量子啟發式算法QUBO中的瓶頸，進而使用JAX框架，將算法移植到GPU上。相較於傳統的C++ CUDA Kernel，Cublas，JAX提供了一種更為簡便的擴展方式來適應我們的算法。這種簡便性不僅體現在程式碼的編寫上，更在於其對於算法調整的高效率和靈活性，

「我們體驗到『奔跑吧，不要用走的』，過程我們實現了高達126倍的End-to-end加速效果，這一結果遠遠超出了我們的預期！」 — 長庚大學資工系Schro?dinger's Cat團隊



更多資訊請看：https://github.com/nqobu/nvidia/raw/main/20231207/Team01.pdf

Mentored by NVIDIA Reese Wang, Yun-Yuan Wang, CK Lee.

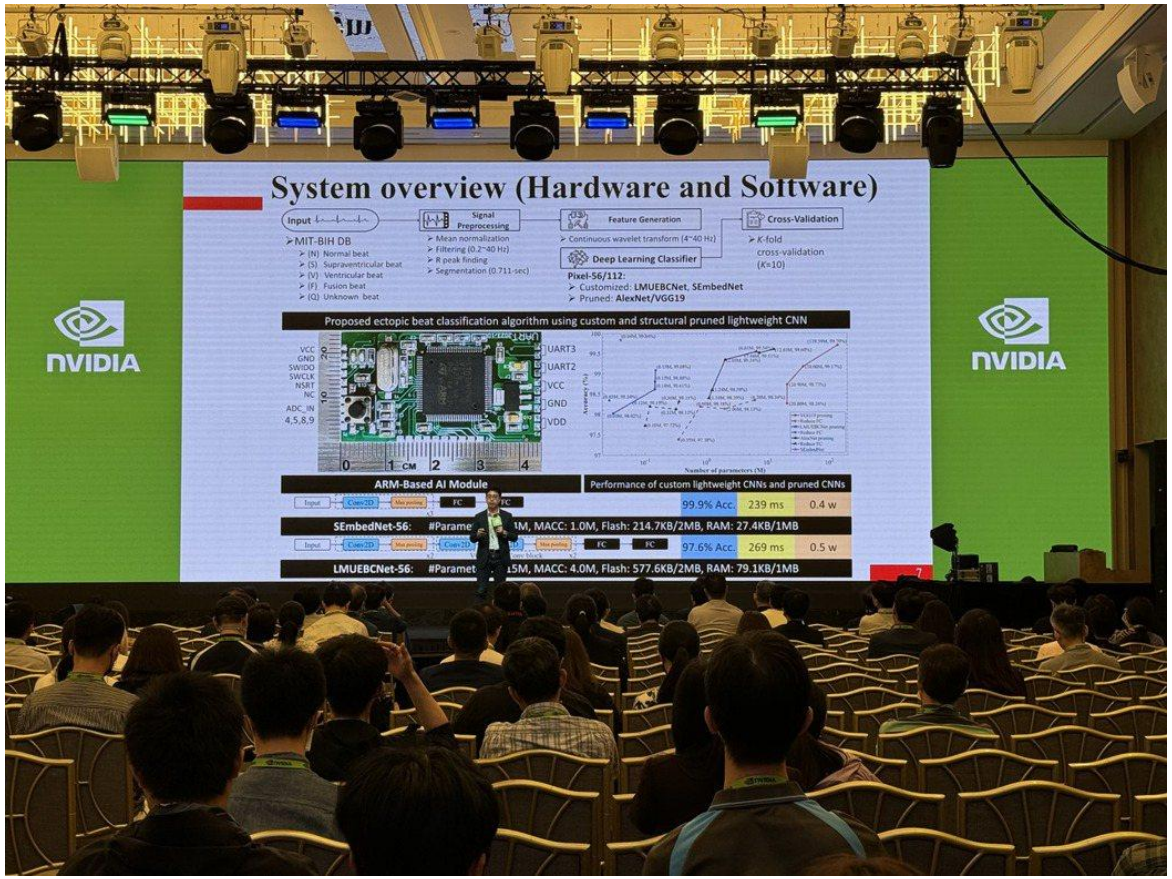
(extend from 2023 Open Hackathon)

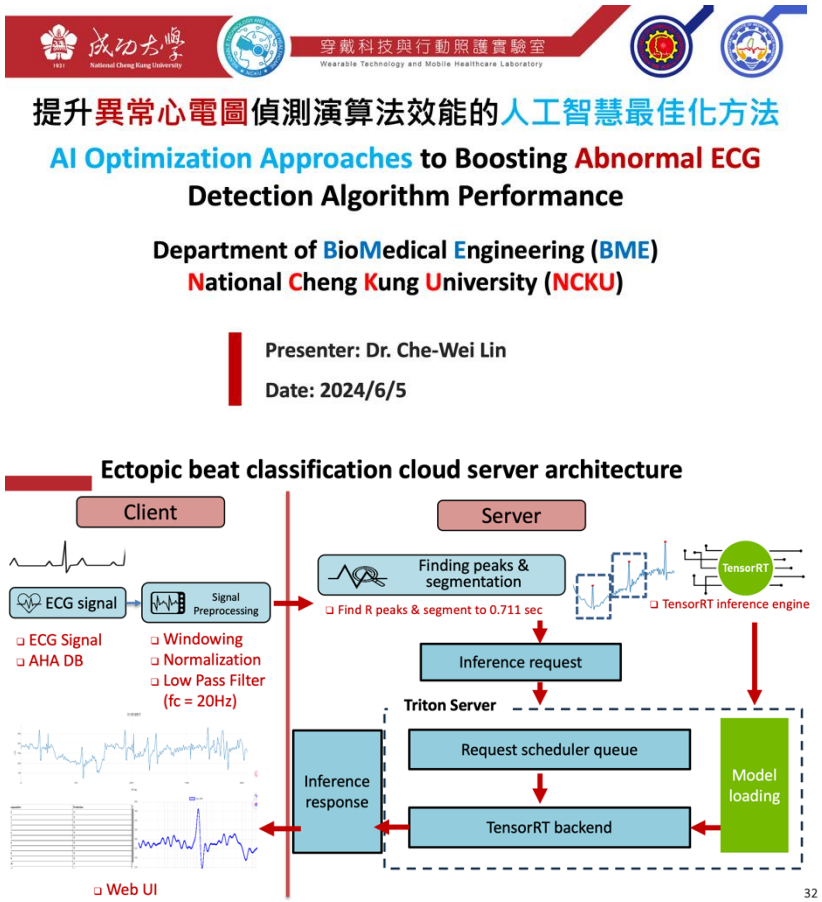
Computex 2024 AI Summit Talk

Prof. Che-Wei Lin @ NCKU-BME

https://www.nvidia.com/en-us/on-demand/session/aisummit24-stc4/

https://github.com/nqobu/nvidia/raw/main/20231207/Team08.pdf





Mentored by NVIDIA Ken Liao

Certification

https://github.com/nqobu/nvidia/blob/main/20231207/Certificate_Of_Attendance.pdf

Opportunties to publish at NCHC's website

NARLabs 財團法人國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing

::: | | | EN | A- A A+ |

核心服務 創新技術 科研成果 動態資訊 關於我們

看 OPEN 黑客松如何帶領了技術變革? DPU把網路，GPU把大型語言模型、大氣科學、量子電路模擬，通通加速！

2023.12.07

動態資訊

- 焦點新聞
- 活動專區
- 動態集錦
- 招標公告
- 出版品專區
- 專利/商標與技術移轉
- 影音專區
- 輿情回應

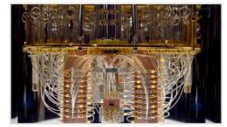
NARLabs 財團法人國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing

::: > 科研成果 > 學研成果

學研成果

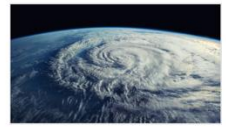
日期 開始日期 ~ 結束日期 關鍵字 關鍵字

ex：2019/01/01



2024.01.30

【2023 NCHC, NVIDIA, OpenACC 黑客松】-HPC、DPU及量子運算加速成果



2024.01.30

【2023 NCHC, NVIDIA, OpenACC 黑客松】-大氣科學應用加速成果

2024.01.30

【2023 NCHC, NVIDIA, OpenACC 黑客松】-人工智慧應用加速成果

OpenACC More Science, Less Programming

OPEN HACKATHONS

NARLabs 財團法人國家實驗研究院 國家高速網路與計算中心 National Center for High-performance Computing

NVIDIA

淺顯易懂的標題

研究領域示意圖

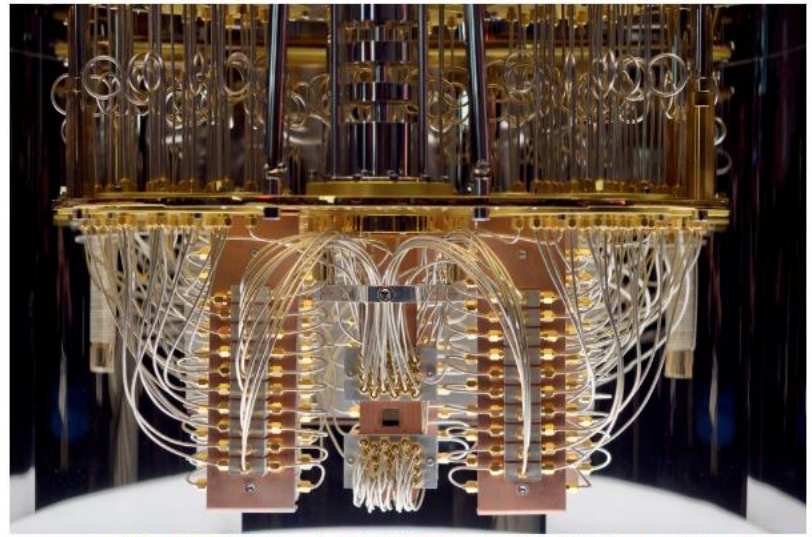
{__} 團隊來自 {__}老師帶領的{__}實驗室，將 {__} 加速了 {__} 倍！！

250~500 字儘量易懂的描述。
為什麼這個領域重要？為什麼這個命題重要？
加速成果？加速帶來什麼影響和潛力？

和加速成果相關的補充數據

報告投影片連結 (由國網上傳到 github)

量子算法模擬



haofan2023團隊成員來自臺灣大學資工系「洪士灝老師實驗室」，將量子演算法QAOA加速468倍！
— NVIDIA Mentors: Tian Zheng, Frank Lin, Yun-Yuan Wang

量子技術正以驚人的速度發展，預示著我們即將進入量子計算的時代。在這個過程中，量子電路模擬成為一個關鍵工具，它在量子硬體和軟體的開發中扮演著重要的角色，特別是在處理量子程式的編寫和驗證方面。傳統電腦的強模擬能夠獲得完整的量子狀態信息。這使得傳統電腦在構建量子系統方面變得不可或缺，尤其是在當前噪聲較多的中等規模量子（NISQ）時代。

量子近似優化算法（QAOA）是一種常用的量子算法，用於通過近似解來解決組合優化問題。然而，在虛擬量子計算機上執行QAOA對於解決需要大規模量子電路模擬的組合優化問題而言，會遇到模擬速度較慢的問題。團隊使用數學優化來壓縮量子操作，並結合有效的位元操作進一步降低計算複雜性，透過GPU加速最高獲取468倍的加速效果！

Table 1: The elapsed time of 5-level QAOA (unit: second, double).

Qubit	CPU_{Single}	$CPU_{Mutiple}$	CPU_{Cache}	GPU_{Cache}	GPU_{All}
23	29.80	1.28 (23x)	1.28 (63x)	0.24 (120x)	0.06 (**341x**)
24	68.00	3.46 (20x)	3.46 (43x)	0.55 (123x)	0.12 (**382x**)
25	152.52	15.32 (10x)	15.31 (45x)	1.19 (127x)	0.23 (**404x**)
26	330.69	33.83 (10x)	33.83 (56x)	2.60 (126x)	0.56 (**417x**)
27	712.26	72.66 (10x)	72.66 (54x)	5.59 (127x)	1.08 (**427x**)
28	1556.87	156.52 (10x)	156.52 (54x)	11.96 (130x)	2.17 (**445x**)
29	3325.55	335.09 (10x)	335.09 (49x)	25.73 (129x)	4.45 (**451x**)
30	7226.46	718.33 (10x)	718.33 (47x)	55.20 (130x)	9.22 (**468x**)

更多資訊請看：https://github.com/nqobu/nvidia/raw/main/20231207/Team02.pdf

Opportunties to publish by media interview

智慧製造的關鍵項目，產線智慧排程如何優化？

由林群惟博士、蘇粲程博士所帶領的團隊「AI Scheduler」，首度參與本屆黑客松，就透過 GPU 加速找到未來商品優化的方向。

因應彈性製造與生產時代來臨，台灣高科技製造業、生技產業等都面臨要快速排程生產的挑戰。這時候透過數位工具來做智慧排程處理，可以有效提升效率與客戶滿意度。

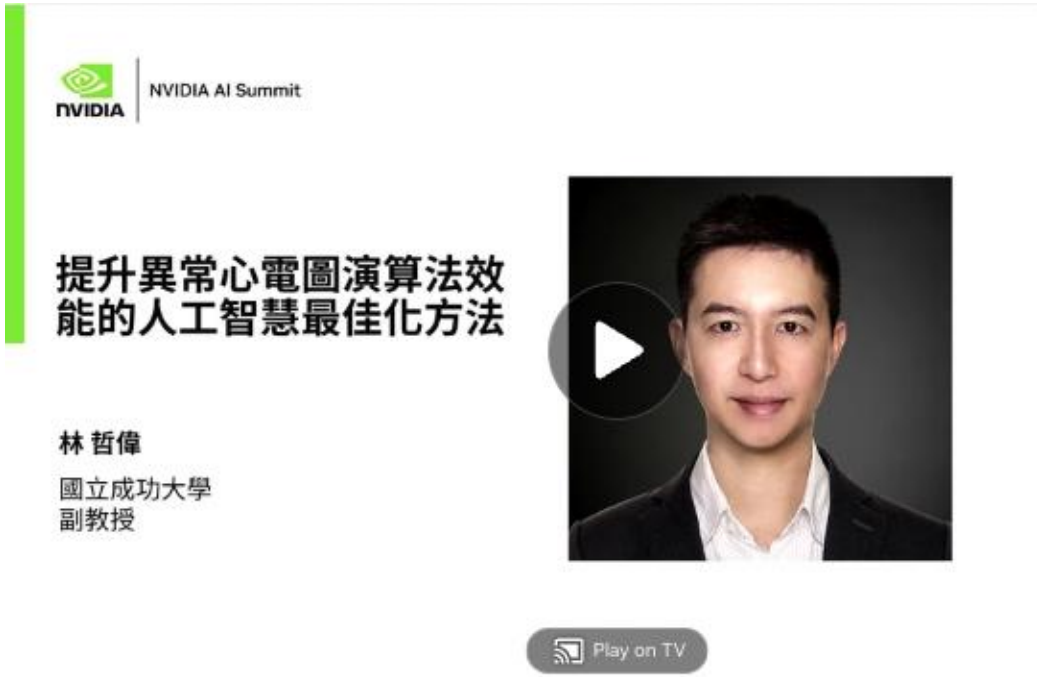
GPU 加速深化台灣地球科學研究成果，提升全球學術圈重要性

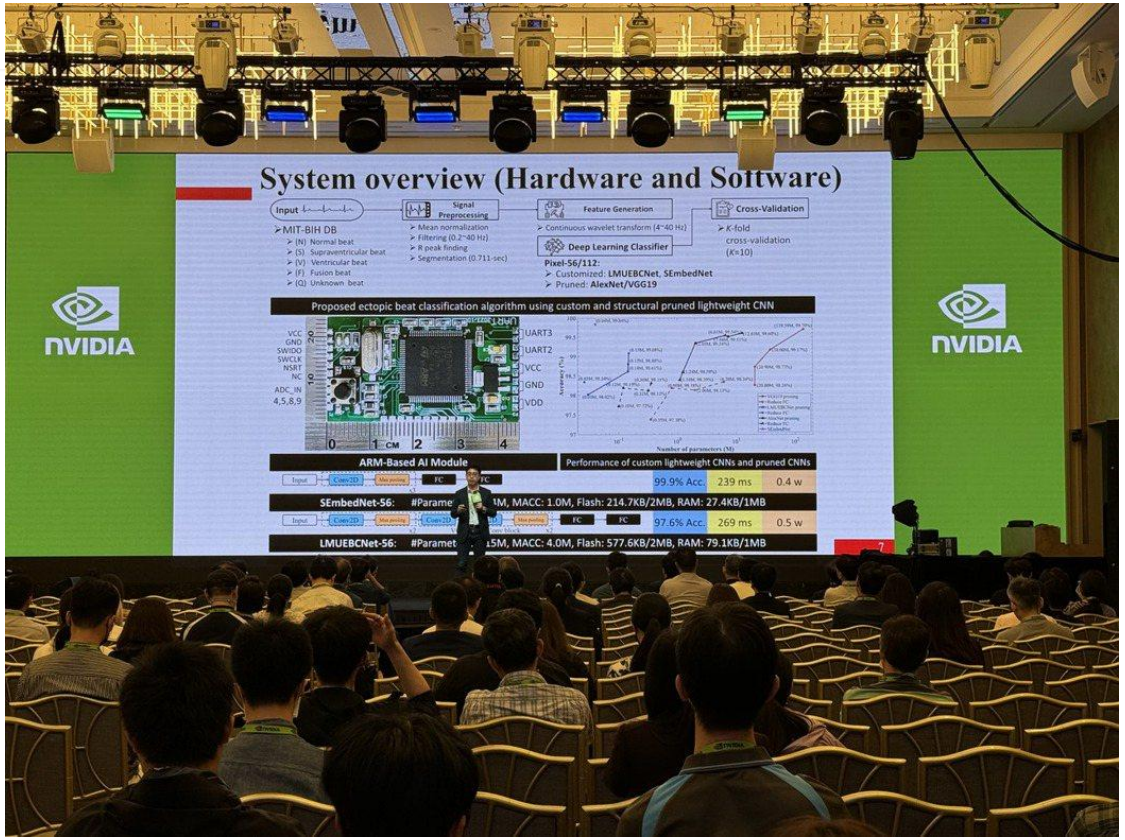
譚老師團隊照片，譚謂（圖中）帶領團隊參與本次 NVIDIA 黑客松，找到地質研究的運算新方法。（圖片：譚謂團隊提供）

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

Opportunties to give a talk in Computex 25 NVIDIA AI Summit

https://www.nvidia.com/en-us/on-demand/session/aisummit24-stc4/

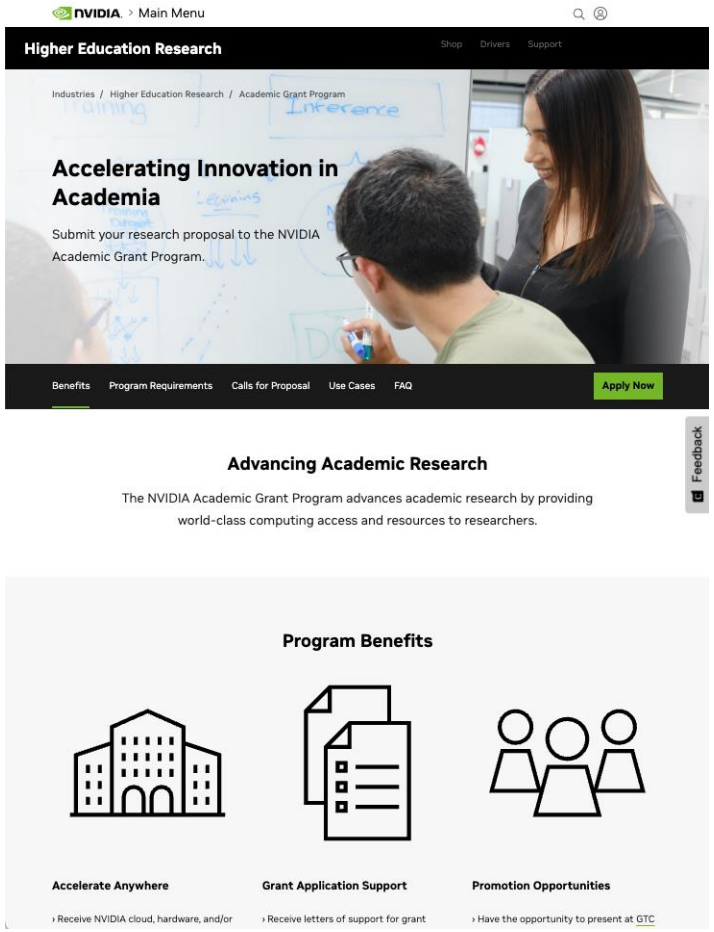




Mentored by NVIDIA Ken Liao

Opportunities to Apply the NVIDIA Academic Grant Program

https://www.nvidia.com/en-us/industries/higher-education-research/academic-grant-program/



Enhancing Vertical-Axis Wind Turbine Performance in Turbulent Flow Using the Direct Forcing Immersed Boundary Method

Principal Investigator Information
Name Title/Role: Ming-Jyh Chern / Professor of Mechanical Engineering.
University Name: National Taiwan University of Science and Technology.

Project Collaborators
1. Name Title/Role: Fandi D. Suprianto / Ph.D. student.
   University Name: National Taiwan University of Science and Technology.
2. Name Title/Role: Desta Goytom Tewolde / Postdoctoral Researcher.
   University Name: National Taiwan University of Science and Technology.

NVIDIA Contact
Mr. Jay Chen, Data Scientist. Email: jaych@nvidia.com
Mr. Jay Chen was one of our mentors in an Open Hackathon hosted by the National Center for High-Performance Computing (NCHC), Taiwan.

Abstract
This project explores the use of the Immersed Boundary (IB) technique coupled with Large Eddy Simulation (LES) to model fluid-structure interactions in Magnus effect and Savonius VAWTs. The IB approach, which uses a fixed Cartesian grid, avoids the challenges of rotating and stationary mesh zones, allowing for more efficient simulations of complex turbine dynamics. However, the high computational demand due to the need for fine grid resolutions, especially near the immersed boundaries, presents a challenge. To address this, we have optimized our in-house Direct Forcing Immersed Boundary method combined Large Eddy Simulation (DFIB-LES) solver, enhancing its efficiency by offloading computations to GPUs. This enables cost-effective and detailed simulations of dynamic rotors under various conditions, facilitating the investigation of performance enhancement strategies for Magnus effect and Savonius VAWTs through two-way fluid-structural coupling.
Keywords: *Vertical-axis wind turbines, Direct Forcing Immersed Boundary method, Large Eddy Simulation.*

NVIDIA Platforms
Software: NVIDIA HPC SDK (with support for CUDA, OpenACC, and MPI)
This is the compiler that we use either on our lab's servers or on the NCHC's HPC cluster.
Hardware: NVIDIA DGX-A100
We had the opportunity to use this machine while participating in the Open Hackathon, as mentioned in Appendix A.

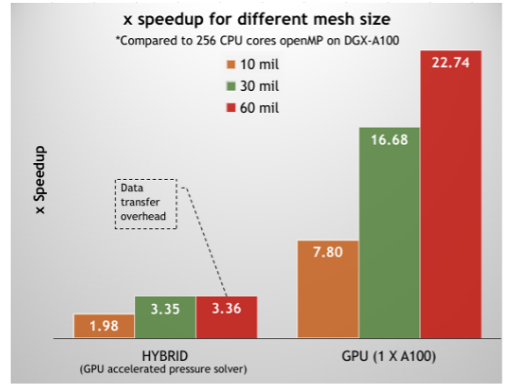
Introduction
Vertical-axis wind turbines (VAWTs) are primarily used for small-scale applications with power ratings of a few kilowatts, making them suitable for both urban and remote areas with less predictable wind speeds. One type of VAWT that has recently gained prominence due to its unique ability to generate electricity even in strong winds is the Magnus effect VAWT. Accurately predicting the aerodynamic performance of VAWTs relies on selecting appropriate computational methods, including turbulence models and numerical schemes. Conventional modeling of VAWTs using a body-fitted grid necessitates stationary and rotating mesh zones. However, this approach can encounter stability issues due to mesh degradation, prolonged data transfer times between zones, and the risk of losing critical unsteady simulation data during transitions between these zones, particularly in complex geometries like VAWT rotors. The Immersed Boundary (IB) technique, on the other hand, employs a fixed Cartesian grid to effectively model complex fluid-structure interactions (FSI), making it an attractive option for simulating VAWTs. However, when coupled with the

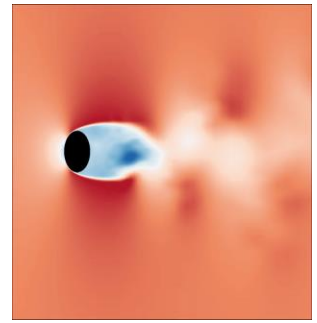
NTUST CFD LAB團隊成員來自臺灣科技大學機械工程系「陳明志老師實驗室」，將3維流體模擬加速16.7倍！
— NVIDIA Mentors: Bharat Kumar, Shijie Wang.

計算流體動力學（CFD）領域正經歷顯著變革，不斷有研究將GPU加速融入其中。這轉變滿足了模擬精度和效率不斷提升的需求，對於複雜工程應用至為關鍵。各種CFD應用程式透過GPU加速展現顯著效能提升，為科學和工程帶來新的可能性。

臺科大團隊使用直接施力沈浸邊界方法（Direct Forcing Immersed Boundary, DFIB）進行三維計算流體力學（3D-CFD），搭配大渦模擬（Large eddy simulation, LES）紊流模型。針對流固耦合（Fluid-structure interaction, FSI）問題尋求解決方案。

採用OpenACC搭配NVTX分析，將主迭代中的所有計算過程轉移到GPU上實現了16.7倍的加速。





「未來研究需要大規模的計算來模擬現實的流體情況，這次取得GPU計算顯著地加速成果，能使我們更進一步實現該目標」— 臺科大機械系NTUST CFD LAB團隊

更多資訊請看：https://github.com/nqobu/nvidia/raw/main/20231207/Team04.pdf

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

Opportunties to Access Taipei-1 DGX H100s

Only scalable multi-node research projects are eligible.

OpenACC More Science, Less Programming

OPEN HACKATHONS

NCHC

NARLabs 財團法人國家實驗研究院 國家高速網路與計算中心 National Center for High-performance Computing

NVIDIA

Comprehensive coverage across DPU, Audio, LLM, Chemistry, Healthcare, Robotics, Weather, Quantum.

	HPC	AI	Quantum	SUM
DPU	清大 周志遠老師			1
Audio		北科大 陳詩雯老師		1
LLM		台大 郭彥甫老師 交大 蔡孟勳老師		2
Chemistry			台大 管希盛老師 台大&中原 張慶瑞老師	2
Healthcare	國網& 長庚 高博	成大 詹寶珠老師		2
Navigation/Robotics	清大 周志遠老師	台&清大 李濬屹老師		2
Weather	台大 吳建銘老師 氣象署&興大 鄧家豪&陳律閎老師			2
SUM	5	5	2	12

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

Kick-off Meeting (Online)

2024 / 11 / 13 (Wed) ： Day 0: Kick-off Meeting 14:00~16:00 PM (Online)

- 02:00 PM - 02:05 PM: Welcome & Opening
- 02:00 PM - 02:15 PM: Event Overview
- **02:15 PM - 03:15 PM: Team self-introduction and getting to know the mentor.**
  - 1 min for all mentors per team
  - 3 mins for each team lead
- 03:15 PM - 03:30 PM: Introduction to computing resources
- 03:30 PM - 04:00 PM: Team discussion
  - **Before** the event, Please self-study the profiling training materials, and see more here.
  - **During** the event, Team discussion on profiling techniques. (Each team's mentors will create a meeting link)
  - **After** the event, Post general profiling questions in #profiler-support.

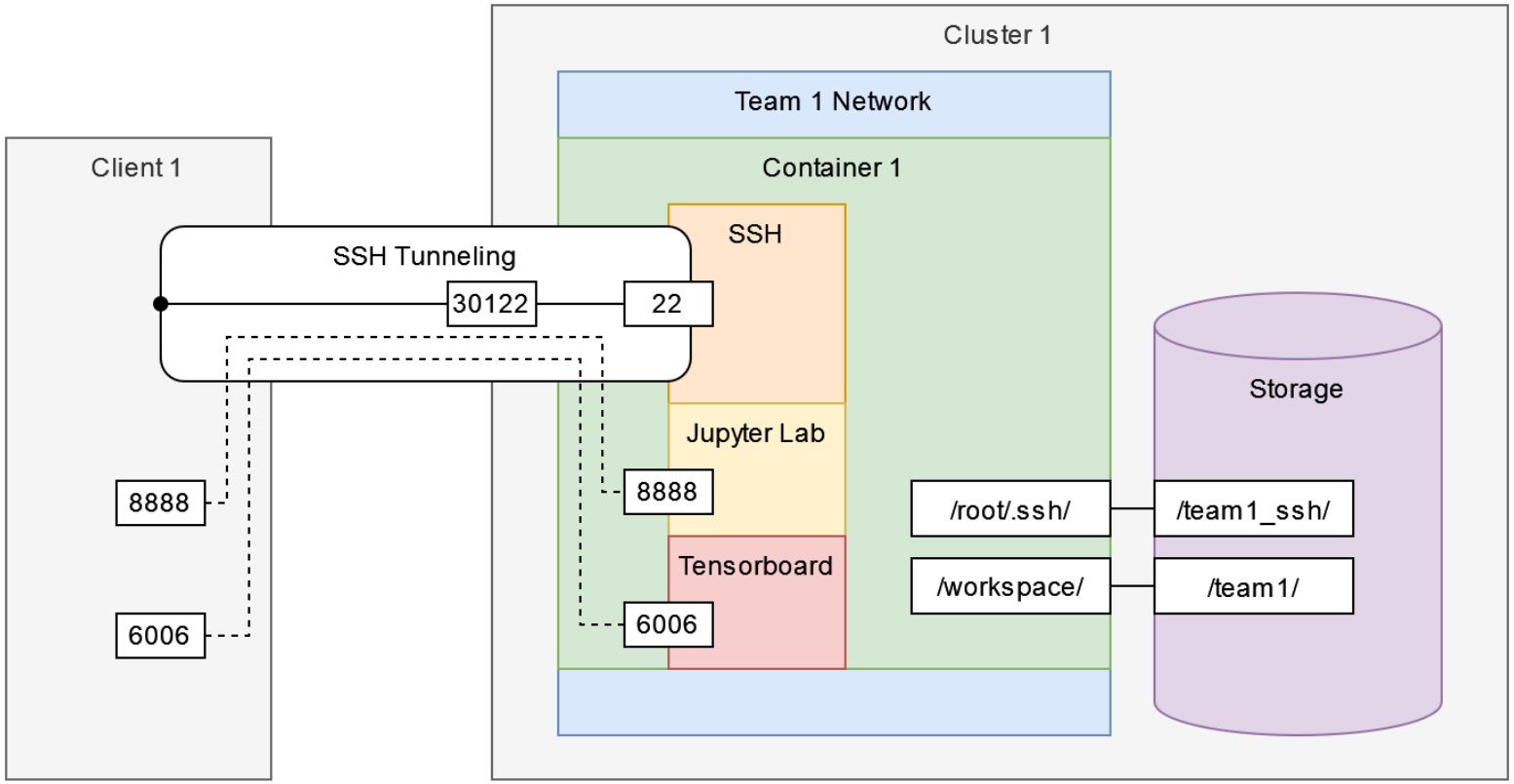
Team self-introduction and getting to know the mentor.

1 min for all mentors per team
3 mins for each team lead

- #team-1-dream-chaser
- #team-2-nycu-hpc-team2
- #team-3-氣象署-興大應數聯隊
- #team-4-ntut_birdsong
- #team-5-parallel-minds
- #team-6-nthu_lsalab
- #team-7-nolab
- #team-8-elsa-robotics
- #team-9-gba-vvm
- #team-10-smile-lab
- #team-11-plantmen
- #team-12-quantum-walk

Access to Machine

Visit Slack Channel #cluster-support




```
ssh root@<YOUR_IP> -p <YOUR_PORT> \
    -L 8888:localhost:8888 \
    -L 6006:localhost:6006
```

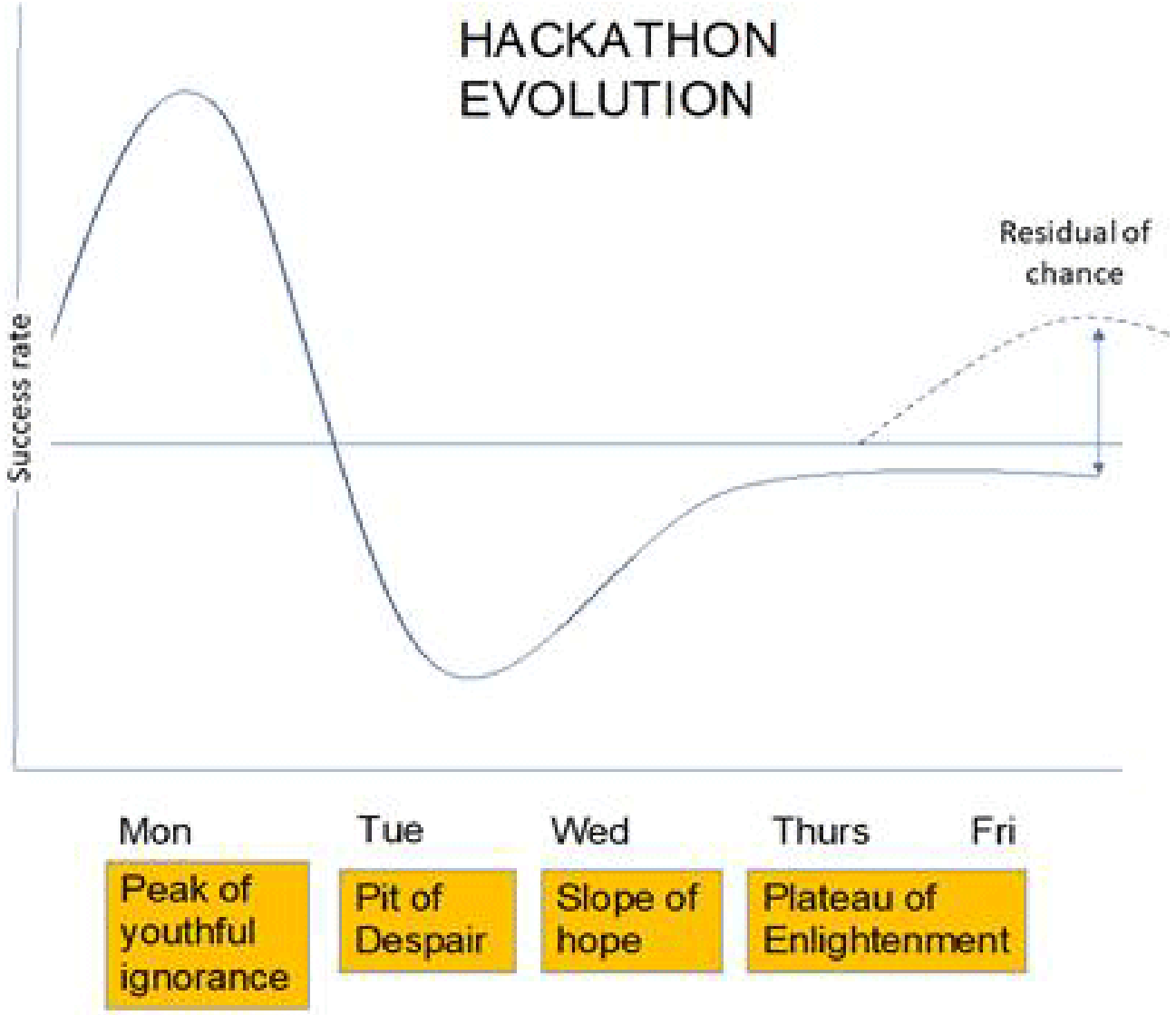
2024 NCHC Open Hackathon

Important Dates

- 2024 / 10 / 15 (Tue)：Registration deadline
- 2024 / 11 / 04~08：Connecting Mentors and Teams
  - Slack Channel, Emails Group, Line Group
- 2024 / 11 / 13 (Wed)：Day 0: Kick-off Meeting 14:00~17:00 PM (Online)
- 2024 / 11 / 20 (Wed)：Day 1: Scrum #1 Meeting 14:00~15:00 PM (Online)
  - 5-minute presentation per team
- 2024 / 11 / 27 (Wed)：Day 2: Scrum #2 Meeting 14:00~15:00 PM (Online)
  - 5-minute presentation per team
- 2024 / 12 / 04 (Wed)：Final Day 10:00~16:00 PM (In-person at NCHC)
  - 12-minute presentation + 3-minute Q&A per team

Profiling Leatning Materials

- [2023 Hackathon Nsight System Session](https://youtu.be/pz69bwhjm3o)
(recommended study this one before the hackathon kick-off formally)
- Learn [NVIDIA profiling tools](https://www.nvidia.com/en-us/on-demand/session/gtcsj20-s22141/) before coming to the event, and visit the "Tools" section on the [Technical Resources page](https://www.openhackathons.org/s/technical-resources).
- lectures in Mandarin
  - [2024/10/24-25 NCHC N-WAY GPU Bootcamp (with 2 Recordings)](https://github.com/nqobu/nvidia/tree/main/20240924)
  - [2023 Hackathon Nsight System Session](https://youtu.be/pz69bwhjm3o)
- lectures in English
  - [Basic Languages Tutorial given by Jeff Larkin/NVIDIA](https://drive.google.com/file/d/1phZGnFnss6iYtJrHYUsKRpwQNB63mLAU/view?usp=sharing)
  - [Math Library: cuSparse and cuSolver Overview by Samuel Rodriguez / Federico Busato](https://drive.google.com/file/d/1EgUYtgpqCr51jC9WWUz2W5wqx-k3aS0s/view?usp=sharing)
  - [Nsight Tutorial given by Max Katz/NVIDIA](https://drive.google.com/file/d/1TEPiRpxqZXK2iqzy1uAQoAlrH3u7z-iX/view?usp=sharing)
  - [Nsight Compute + Nsight System Q&A + Demo by Chris Ashton / Jackson Marusarz / Tod Courtney](https://drive.google.com/file/d/1nt6cfGNfFU-jdj6ZVYmjOT_DQx5m-JVL/view?usp=sharing)
  - [Profiling with TensorFlow Demo, presented by Kaleb Smith/NVIDIA](https://drive.google.com/file/d/1DXO0xNZj_zrN46HszCyPX-BJMqFQ_Er5/view?usp=sharing)
  - [Pytorch Nsight System Demo given by Tod Courtney/NVIDIA](https://drive.google.com/file/d/1EAfSgt1UzPRgP_xkf3zDnVvdumfnDk_2/view?usp=sharing)



This article highlights the Oak Ridge Leadership Compute Facilitys GPU Hackathon
https://www.computer.org/csdl/magazine/cs/2018/04/mcs2018040095/13rRUxYIN88

Day 0 - Template

OpenACC
More Science, Less Programming

Total presentation time is 3 minutes

Team Name

Team Members (Name, organization, and picture)
Team Mentors (Name, organization, and picture)

Your Code/Application

- Tell us about your application:
  - What's the algorithmic motif?
  - Libraries?
  - Language?
  - Which application module/function are you focusing on?
  - GPU port path (CUDA/OpenACC/OpenMP)

Goals

- What would you like to achieve by the end of the week?

Day 1&2 - Template

OpenACC
More Science, Less Programming

Total presentation time is 4 minutes

Team Name

Team Members (Name, organization, and picture)
Team Mentors (Name, organization, and picture)

Progress and Goals

- What have you accomplished since yesterday?
- What are your goals for today?

Profiler Output

Problems and Solutions

- What problems are you currently facing?
- Have you resolved any problems (or found bugs) that others might find useful?

Day Final - Template

OpenACC
More Science, Less Programming

Total presentation time is 12 minutes

Team Name

Team Members (Name, organization, and picture)
Team Mentors (Name, organization, and picture)

Application Name

- Problem the team is trying to solve.
- Scientific driver for the chosen algorithm.
- What's the algorithmic motif?
- What parts are you focused on?

Evolution and Strategy

- What was your goal for coming here?
- What was your initial strategy?
- How did this strategy change?

Results and Final Profile

- What were you able to accomplish?
- Did you achieve a speed up?
- Show multi-core vs. GPU numbers
- What did you learn?
- Did you create a new algorithm?
- Did you achieve new scientific goals?

Energy Efficiency

INPUTS	
# CPU Cores	64
# GPUs (A100)	6
Application Speedup	20.0x

Node Replacement	13.3x

GPU NODE POWER SAVINGS			
	AMD Dual Rome 7742	8x A100 80GB SXM4	Power Savings
Compute Power (W)	14,667	6,500	8,167
Networking Power (W)	619	93	526
Total Power (W)	**15,286**	**6,593**	8,693

Node Power efficiency	2.3x

ANNUAL ENERGY SAVINGS PER GPU NODE			
	AMD Dual Rome 7742	8x A100 80GB SXM4	Power Savings
Compute Power (kWh/year)	128,480	56,940	71,540
Networking Power (kWh/year)	5,424	814	4,610
Total Power (kWh/year)	**133,904**	**57,754**	**76,150**

$/kWh	$	0.18
Annual Cost Savings	$	13,707.04
3-year Cost Savings	$	41,121.13

Metric Tons of CO2	54
Gasoline Cars Driven for 1 year	12
Seedlings Trees grown for 10 years	892

(source: Link)

The calculator will compare the consumption of a number of CPU-only nodes with dual CPUs required to perform the same amount of work as 1 GPU node with 2 CPUs and 8 GPUs.

1. Use this calculator for your report
2. Add your acceleration numbers in the INPUTS section
3. Modify $/kwh number if necessary
4. Paste a screenshot similar to the one on the right in this slide to report energy efficiency of your project

What problems have you encountered?

- Problems with legacy app structure.
- Issues with algorithms.
- Tool bugs.
- Tool lacking features.
- System setup.

Wishlist

- What do you wish existed to make your life easier?
- Tools
- Language standards
- Event
- Systems

Final Thoughts

- Was this Open Hackathon worth it?
- Will you continue development?
- Next steps, future plans.
- What sustained resources or support will be critical for your work after the event?

Application Background

High-level description of application and uses.
Light on domain-specific jargon; should be appropriate for a general technical audience.
Targeted computational motifs.

Plot Showing Representative Result

Caption describing figure in simple terms

Hackathon Objectives and Approach

Programming models.
Profiling/hot spots
Refactoring
Libraries
Performance tuning
Other

Technical Accomplishments and Impact

What were you able to achieve at the hackathon?
How did you achieve it?
Speedup
Why does it matter/what does it enable?

Opportunties to publish at NCHC's website

NARLabs 財團法人國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing

::: | | | EN | A- A A+ |

核心服務 創新技術 科研成果 動態資訊 關於我們

看 OPEN 黑客松如何帶領了技術變革? DPU把網路，GPU把大型語言模型、大氣科學、量子電路模擬，通通加速！

2023.12.07

動態資訊

- 焦點新聞
- 活動專區
- 動態集錦
- 招標公告
- 出版品專區
- 專利/商標與技術移轉
- 影音專區
- 輿情回應

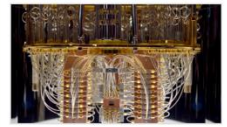
NARLabs 財團法人國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing

::: > 科研成果 > 學研成果

學研成果

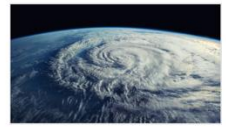
日期 開始日期 ~ 結束日期 關鍵字 關鍵字

ex：2019/01/01



2024.01.30

【2023 NCHC, NVIDIA, OpenACC 黑客松】-HPC、DPU及量子運算加速成果



2024.01.30

【2023 NCHC, NVIDIA, OpenACC 黑客松】-大氣科學應用加速成果

2024.01.30

【2023 NCHC, NVIDIA, OpenACC 黑客松】-人工智慧應用加速成果

OpenACC More Science, Less Programming

OPEN HACKATHONS

NARLabs 財團法人國家實驗研究院 國家高速網路與計算中心 National Center for High-performance Computing

NVIDIA

Opportunties to publish at NCHC's website

Please summarize your team's achievements during the Open Hackathon (see below example).

淺顯易懂的標題

研究領域示意圖

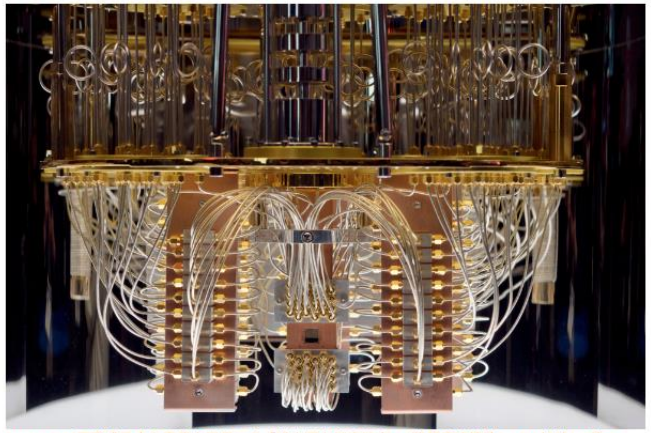
{__} 團隊來自 {__}老師帶領的{__}實驗室,
將 {__} 加速了 {__} 倍！！

250~500 字儘量老嫗能解的描述。
為什麼這個領域重要？為什麼這個命題重要？加速成果？加速帶來什麼影響和潛力？

和加速成果相關的補充數據

報告投影片連結 (由國網上傳到 github)

量子算法模擬



haofan2023團隊成員來自臺灣大學資工系「洪士灝老師實驗室」，將量子演算法QAOA加速468倍！

— NVIDIA Mentors: Tian Zheng, Frank Lin, Yun-Yuan Wang

量子技術正以驚人的速度發展，預示著我們即將進入量子計算的時代。在這個過程中，量子電路模擬成為一個關鍵工具，它在量子硬體和軟體的開發中扮演著重要的角色，特別是在處理量子程式的編寫和驗證方面。傳統電腦的強模擬能夠獲得完整的量子狀態信息。這使得傳統電腦在構建量子系統方面變得不可或缺，尤其是在當前噪聲較多的中等規模量子（NISQ）時代。

量子近似優化算法（QAOA）是一種常用的量子算法，用於通過近似解來解決組合優化問題。然而，在虛擬量子計算機上執行QAOA對於解決需要大規模量子電路模擬的組合優化問題而言，會遇到模擬速度較慢的問題。團隊使用數學優化來壓縮量子操作，並結合有效的位元操作進一步降低計算複雜性，透過GPU加速最高獲取468倍的加速效果！

Table 1: The elapsed time of 5-level QAOA (unit: second, double).

Qubit	CPU_{Single}	$CPU_{Mutiple}$	CPU_{Cache}	GPU_{Cache}	GPU_{All}
23	29.80	1.28 (23x)	1.28 (63x)	0.24 (120x)	0.06 (**341x**)
24	68.00	3.46 (20x)	3.46 (43x)	0.55 (123x)	0.12 (**382x**)
25	152.52	15.32 (10x)	15.31 (45x)	1.19 (127x)	0.23 (**404x**)
26	330.69	33.83 (10x)	33.83 (56x)	2.60 (126x)	0.56 (**417x**)
27	712.26	72.66 (10x)	72.66 (54x)	5.59 (127x)	1.08 (**427x**)
28	1556.87	156.52 (10x)	156.52 (54x)	11.96 (130x)	2.17 (**445x**)
29	3325.55	335.09 (10x)	335.09 (49x)	25.73 (129x)	4.45 (**451x**)
30	7226.46	718.33 (10x)	718.33 (47x)	55.20 (130x)	9.22 (**468x**)

更多資訊請看：https://github.com/nqobu/nvidia/raw/main/20231207/Team02.pdf

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

2023 Final Report Examples

https://github.com/nqobu/nvidia/tree/main/20231207

參與團隊

1. Quantum-Inspired Algorithm (QUBO), Schrödinger's cat
2. Quantum Circuit Simulation (QFT, QAOA), haofan2023
3. 5G SBA (Service Based Architecture), NTHU-LSALAB
4. 3D-CFD (Direct Forcing Immersed Boundary, LES turbulence model), NTUST CFD Lab
5. Mesh Generation for MPAS Model, CWA mesh generation
6. A.I. in Otoscopic Diagnosis, CYCU BME
7. Global Ensemble model Verification, CWA_GVER
8. Arrhythmia Screening of Real-Time Single-Lead ECG, WTMH
9. CWAGFS-TCo - Numerical Weather Prediction Model, 氣象署-興大應數聯隊
10. Accelerate Encrypt/Decrypt Operation in Functional Encryption, YSS Team
11. X-ray Background Correction Model, TXM AI Group
12. LLM Inference with TensorRT-LLM on NCHC servers, NCHC Speedrunning Team


Thank You
OpenACC
More Science, Less Programming