

ASIANSCIENTIST

Issue 15  
January 2024

# SUPERCOMPUTING ASIA

## AN EYE ON AI

5 MAJOR BREAKTHROUGHS  
FUELED BY HPC



ACCELERATING ADVANCEMENTS  
IN HEALTHCARE

DEFINING THE FUTURE  
OF AI

THE KEY TO CLIMATE  
MODELS



SCA 2024  
Supercomputing Asia  
Gathering the Best of HPC in Asia

TPC

# SupercomputingAsia 2024

Exascale readiness in AI, HPC, and Quantum



19-22 February 2024



International  
Convention Centre  
Sydney, Australia

11

KEYNOTE  
TALKS

7

TECHNICAL  
TALKS  
115

WORKSHOPS

SPONSORS

50+

## ACTIVITIES INCLUDE :

- ✓ Doctoral Showcase Competition
- ✓ ED&I Panel
- ✓ Best Paper/Poster Awards
- ✓ Career Fair Session

 [sca24.sc-asia.org](https://sca24.sc-asia.org)

 [SCAsia2024@gmail.com](mailto:SCAsia2024@gmail.com)



# CONTENTS

Issue 15  
January 2024

## FEATURES

p. 24

### DEFINING THE FUTURE OF AI

*A chat with  
Torsten Hoefer of  
the Swiss National  
Supercomputing  
Centre*

## COVER STORY

p. 16

### An Eye On AI: 5 Major Breakthroughs Fueled By HPC

*An overview of how HPC has  
transformed AI, and how AI  
has transformed our world*

## FEATURES

p. 10

### Accelerating Advancements In Healthcare

*Healthcare gets a boost from  
AI and HPC*

p. 30

### The Key To Climate Models

*Supercomputing deals with  
a deluge of climate data*

p. 06

### Digital Dispatch

*Supercomputing news from  
across the region*

p. 36

### Business Bytes

*The latest industry moves*

p. 38

### Super Snapshot

*Charting a research roadmap  
with Quek Gim Pew*

# EDITOR'S NOTE

Since the early 2000s, experts have speculated that—just like the steam engine and the internet—artificial intelligence (AI) would change the world. With the AI revolution well on its way, supercomputers play a significant role in the journey as they process swathes of data and support the development of AI across a wide range of applications.

In this issue of *Supercomputing Asia*, we cover AI's presence in two key areas: healthcare and the climate. From drug development to disease diagnostics, healthcare systems all over the world stand to be improved by the power of high performance computing (HPC) and AI (*Accelerating Advancements in Healthcare*, p. 10). To inform climate policies and infrastructure plans, governments look to climate modeling—a field that aims to get more precise each year with the help of HPC (*The Key to Climate Models*, p. 30).

While HPC supports AI advancements in large language models, drug discovery and climate prediction, AI in turn supports the development of faster and more efficient supercomputers. For a bird's eye view of the relationship between supercomputers and AI, our cover story highlights five major AI breakthroughs fueled by HPC (*An Eye on AI*, p. 16).

Finally, we consider the future of AI with *Supercomputing Asia's* first interview profile (*Defining the Future of AI*, p. 24). Professor Torsten Hoeller, Chief Architect for Machine Learning at the Swiss National Supercomputing Centre, discusses major opportunities and challenges in the space as well as the importance of both collaboration and education.

**Juliana Chan, Ph.D.**  
CEO & Publisher  
*Supercomputing Asia*



- /asianscientist
- @asianscientist
- Asian Scientist Magazine
- asianscientist
- AsianScientist
- asianscientist

[www.asianscientist.com](http://www.asianscientist.com)

## SUPERCOMPUTING ASIA

### EDITORIAL ADVISORY COMMITTEE

Prof. Tan Tin Wee  
Prof. Satoshi Matsuoka  
Prof. John Gustafson  
Yves Poppe

### CEO & PUBLISHER

Dr. Juliana Chan

### MANAGING EDITOR

Joanne Chow

### EDITORS

Jill Arul  
Rachel Soon

### CONTRIBUTORS

Chia Pei Ling  
Mitchell Lim  
Lee Kai Xiang  
Erinne Ong

### ART DIRECTOR

Shelly Liew

### JUNIOR ART DIRECTOR

Lieu Yi Pei

### SENIOR DESIGNER

Ajun Chuah

### DESIGNER

Wong Wey Wen

### SALES & MARKETING

Samantha Yeap  
Kata Llamas  
Audrey Tan

### PUBLISHED BY

Wildtype Media Group Pte Ltd

### HEAD OFFICE

**Wildtype Media Group Pte Ltd**  
5 Toh Tuck Link  
Singapore 596224  
[hello@wildtype.media](mailto:hello@wildtype.media)  
Website: [www.asianscientist.com/subscribe](http://www.asianscientist.com/subscribe)



Combining savvy communication with technical rigor, Wildtype Media Group is Asia's leading STEM and healthcare media company, spanning digital, social media, video, print, custom publishing and events.  
Brands under Wildtype Media Group include the flagship *Asian Scientist Magazine* and *Supercomputing Asia*, award-winning titles available in print and online.  
[www.wildtype.media](http://www.wildtype.media)

To order reprints or e-prints, or request permission to republish *Supercomputing Asia* content, please contact [hello@wildtype.media](mailto:hello@wildtype.media)

Supercomputing Asia, MCI (P) 018/05/2023, is published by Wildtype Media Group Pte Ltd. Printed by Times Printers Pte Ltd. All rights reserved. No part of this publication is to be reproduced, stored, transmitted, digitally or otherwise, without the prior consent of the publisher. The information contained herein is accurate at time of printing. Changes may have occurred since this magazine went to print. Supercomputing Asia content is provided 'as is' without warranty of any kind. Supercomputing Asia excludes all warranties, either expressed or implied. In no event shall Wildtype Media Group Pte Ltd and its editors be held liable for any damages, loss, injury, or inconvenience, arising, directly or indirectly, in connection with the contents of the magazine. Any persons relying on Supercomputing Asia content shall do so at their own risk. Advertising material submitted by third parties in Supercomputing Asia is the sole responsibility of each individual advertiser. We accept no responsibility for the content of advertising material, including, without limitation, any error, omission or inaccuracy therein.

# HPCIC23

HPC AI INNOVATION CHALLENGE 2023

INNOVATE, ACCELERATE, TRANSFORM

Organised by **Infocomm Media Development Authority (IMDA), AI Singapore** and the **National Supercomputing Centre (NSCC) Singapore**, the annual High Performance Computing (HPC) Artificial Intelligence Innovation Challenge (HPCIC23) offers top innovators access to exclusive high performance computing power to bring their vision to life and accelerate solutions in these four key domains:



Manufacturing, Trade  
& Connectivity



Human Health &  
Potential



Smart Nation &  
Digital Economy



Urban Solutions &  
Sustainability

The challenge aims to:

- ✓ Address complex real-world issues, promoting collaboration and knowledge exchange in the AI and HPC communities.
- ✓ Bring fresh perspectives to existing problems, driving innovative solutions in AI and HPC.
- ✓ Bridge academia and industry, empowering students to build skills and offering companies access to new talent and inventive ideas.

Discover the finalists' journey here

[www.hpcic23.sg](http://www.hpcic23.sg)

Organised by:



## SOUTH KOREA SETS ITS SIGHTS ON EXASCALE SUPERCOMPUTING

According to the *Korea Economic Daily*, the Ministry of Science and ICT of South Korea revealed in May 2023 the national plan to boost the country's dominance in high-performance computing (HPC) developments. The plan outlines the future development of quantum computers, new open-source software, HPC-software as a service and industry-tailored computer resource support.

The plan also demonstrates South Korea's ambition for exascale supercomputing, as the building of the country's seventh supercomputer, which will be capable of exaflop performance, will commence in 2025. The seventh supercomputer will far exceed the country's petaflop-performing sixth supercomputer, which is still being built.

The government intends both the sixth and seventh models for industrial use, specifically replacing prototype production with simulation methods in an effort to help local manufacturers reduce cost.

## ADVANCING QUANTUM PROCESSING WITH A SINGLE-PHOTON LIGHT SOURCE AT ROOM TEMPERATURE

Rare-earth (RE) atoms and ions are typically used to develop photon-transmitting optical fibers for quantum processing. However, because optical fiber-based photon light sources are often made with RE-doped crystalline materials at subzero temperatures, expensive cooling systems are required to maintain the quantum networks built using these fibers.

A research team from Japan has created an amorphous silica optical fiber capable of transmitting photons at room temperature. With a heat-and-pull method, the team fabricated a single-photon light source using optical fibers doped with ytterbium ions, a type of RE element that has conducive optical and electronic properties.

Published on October 16, 2023 in *Physical Review Applied*, the research findings give rise to a potential future of cost-effective and accessible quantum networks that do not require cooling systems.

## WINNERS OF SUPERCON 2023 ANNOUNCED

The 2023 Supercomputing Contest (SuperCon 2023) for high school and technical college students in Japan was hosted online from August 21 to 25 by Tokyo Tech's Global Scientific Information and Computing Center, Osaka University's Cybermedia Center and the RIKEN Center for Computational Science. In the 29<sup>th</sup> edition of this competition, three teams emerged victorious:

1<sup>ST</sup>  
PLACE

**Team KMB76**  
*Nada Senior  
High School*

2<sup>ND</sup>  
PLACE

**Team honyanya**  
*Kaisei Senior  
High School*

3<sup>RD</sup>  
PLACE

**Team prism**  
*Senior High School  
at Komaba,  
University of Tsukuba*

SuperCon is an annual competition that challenges students to use a supercomputer to solve a given problem accurately and quickly over the course of four days. For this year's challenge, the teams were tasked to "find the closest pair of points distributed in a two-dimensional space from a given 500 million points".

For placing first in the competition, Team KMB76 received the Minister of Education, Culture, Sports, Science and Technology Award and the Academic Award from the Institute of Electronics, Information and Communication Engineers, and Information Processing Society of Japan.

## A RESEARCH PARTNERSHIP TO DEVELOP CUTTING-EDGE HPC AND AI TECHNOLOGIES IN INDIA

Advanced Micro Devices (AMD) and the Indian Institute of Science (IISc) in Bengaluru have announced a collaborative research initiative focused on advancing high-performance computing (HPC) and artificial intelligence (AI) in India. According to an article published in *Business World* on November 7, 2023, AMD and IISc aim to explore innovations in heterogeneous computing—which include core design, AI training, inference and compilation—and jointly own any intellectual property arising from this collaboration.

In addition, AMD has set up a sponsorship scheme for students pursuing research in low-power designs and machine learning as well as donated a cluster of HPC nodes outfitted with AMD EPYC™ processors, MI Instinct™ graphics processing units (GPUs), and Alveo™ V70 accelerators. The semiconductor giant has also established a research division within the India Development Centre to drive projects in evolving technologies.

## WHAT'S UP!

### SUPERCOMPUTINGASIA 2024

SupercomputingAsia (SCA) returns this year with the theme, “Exascale readiness in HPC, AI, and Quantum”. Taking place at the International Convention Centre Sydney February 19–22, 2024, SCA24 aims to promote a vibrant HPC and AI ecosystem for the public and private sectors in Asia. A major supercomputing conference in Asia, this year’s event is anchored by the National Computational Infrastructure (NCI) in Australia and co-organized by different supercomputing centers across the region, including those in Singapore, Japan, Thailand and New Zealand.

There will be co-located events held in conjunction with SCA24. These include NCI’s 14<sup>th</sup> ADAC Symposium Workshop, which has been included as part of the program of SCA24, as well as AeRO Forum by Australasian eResearch Organisations which invites participants to test the organization’s recently drafted Research Data Reference Architecture with real-world data.

Book your ticket to SCA24 and hear from thought leaders in academia and industry on the latest HPC trends.

For more information, visit  
<https://sca24.sc-asia.org/>

#### WHERE

SYDNEY, AUSTRALIA

#### WHEN

FEBRUARY 19–22, 2024

### ISC HIGH PERFORMANCE 2024

Running on a similar theme as HPC Asia 2024 and SCA24 is this year’s iteration of the International Supercomputing Conference (ISC) High Performance, the annual worldwide gathering of HPC technology providers and users. ISC High Performance 2024 will focus on “reinventing HPC” as Moore’s Law slows down and advancements in HPC, AI and quantum computing continue to evolve and shape the HPC landscape. This theme picks up where the 2023 edition left off, which explored new approaches to HPC.

Like past events, ISC High Performance 2024 will comprise a conference program and an exhibition. The key topics that will be covered at this year’s event range from system architecture and hardware components to machine learning and AI to HPC community engagement and skills development.

ISC High Performance 2024 will be held at the Congress Center Hamburg from May 12 to 16. Don’t miss out on networking and exchanging ideas with vendors, service providers and users of HPC technologies from all over the world.

For more information, visit  
<https://www.isc-hpc.com/>

#### WHERE

HAMBURG, GERMANY

#### WHEN

MAY 12–16, 2024

# ACCELERATING ADVANCEMENTS

## IN HEALTHCARE

From disease diagnostics to drug development, healthcare applications are poised to receive a major boost from the combined power of AI and supercomputing.

By **Erinne Ong**

Photo illustrations by Lieou Yi Pei / *Supercomputing Asia*

**F**or seasoned scientists, it often takes one look at an experimental cohort, whether through microscopic images or a series of electrical waves, to pick apart alterations from the norm. In the clinic, physicians can quickly combine information from a battery of tests to spot signs of disease and deliver an accurate diagnosis.

Their uncanny ability to distinguish healthy from sick can be attributed to years of practice, training and extensive experience with analyzing biological samples.

Inspired by the human mind's capacity for learning, artificial intelligence (AI) models are first trained on existing datasets so that they can recognize patterns and apply the same rules to new samples. This opens avenues for various healthcare applications, like detecting diseases early or making predictions about responses to treatment.

As straightforward as the process may sound, the training phase is typically a gargantuan task, especially considering the variability between individuals and their possible symptoms. High-performance computing (HPC) could be the key to unlocking this bottleneck, offering massive computing power that enables the processing of multitudes of clinical data in a short time span.

By synergizing HPC and AI resources, scientists and physicians can hope to make sense of complex biological phenomena more rapidly and accurately.

## A SHORTER ROUTE TO DRUG DEVELOPMENT

From antibiotic pills to anti-inflammatory ointments, the wonders of modern medicine are perhaps best encapsulated by the spectrum of drugs lining pharmacy shelves. Many previously untreatable diseases can now be addressed by several different therapies, yet many more represent persisting clinical gaps—urgently needing the development of more effective interventions.

But before any drug can be approved for clinical practice, it must first undergo thorough evaluation to demonstrate its medical benefits and outline any potential side effects. This journey of drug discovery and development is often tedious: starting from identifying druggable targets and compounds with possible pharmacological activities, followed by several rounds of testing from cell cultures to pre-clinical models to human clinical trials.

“Drugs are very expensive to develop and the entire process can take 10 to 15 years,” said Professor Satoshi Matsuoka, Director of the RIKEN Center for Computational Science. “One way for costs to go down is by introducing automation and shortening the development cycle.”

Innovations in HPC and AI are in prime position to accelerate the drug development pipeline, without cutting corners nor compromising safety. One of the most critical aspects of synthesizing these compounds lies in performing molecular dynamics simulations, which model atomic motions, interactions and overall conformational changes over time.

Whether through anesthetics that block off pain sensations or carcinogenic agents that trigger several pathways to drive cell proliferation, biomolecules exert their effects primarily by interacting with others. They can have multiple binding sites and various interaction partners, with the nature of such activities changing depending on the molecule’s structure and environmental conditions. Even a small alteration in their structure—and by extension, the genetic code that contains the instructions for producing these molecules—could bring about massive consequences for their functionality.

Thanks to their impressive capacity for running numerous simulations, AI algorithms can help scientists search for candidate drug compounds, discover novel drug targets, delineate their structures, and predict the biochemical interactions between these molecules and the human body. Adding HPC into the mix is akin to shifting into second gear: enabling larger scale, higher quality and much faster simulations to be performed in parallel.

To this end, Taiwan Web Service Corporation (TWSC), a subsidiary of multinational computer hardware company ASUS, has been making significant strides toward building high-precision and seamless workflows for the biomedical sector, backed by HPC and deep learning.

“We have incorporated AI applications into the entire biomedical engineering process to meet the needs of data processing, AI biomedical model training and technology tool creation,” said TWSC CEO Peter Wu in a press release.

By integrating the nine-petaFLOPS Taiwan 2 supercomputer with an optimized GPU framework from NVIDIA, the team is driving the intelligent transformation of various biomedical applications, including bioinformatics analysis and medical imaging.

Forgoing the need for complex programming skills, their OneAI no-code development platform makes secondary gene analysis more easily accessible, enabling users to hunt for potential genomic variants of medical relevance. By leveraging the efficient GPU processing of NVIDIA Parabricks to analyze such complex data, TWSC’s AI supercomputer is 80 times faster than traditional CPU solutions and cuts computational costs in half. The NVIDIA Clara for Drug Discovery deep learning algorithms further bolster these endeavors, performing molecular dynamics simulations and protein structure prediction to accelerate the development of new drugs.

## GENERATING DRUG CANDIDATES

In another collaborative effort, NVIDIA and Japanese corporation Mitsui have joined forces for the Tokyo-1 project, using NVIDIA’s HPC resources for molecular dynamics and generative AI (GenAI) models. The NVIDIA DGX system features dual x86 CPUs and eight H100 Tensor Core GPUs, with each contributing 32 petaFLOPS of computing power to take on massive workloads such as running large language models (LLMs) involving millions of parameters.

LLMs may seem like a misnomer, having surged to popular consciousness particularly through ChatGPT. However, language is not limited to human speech and can also encompass the language of biochemistry. One’s DNA is essentially a string of chemicals that can be represented by a “letter” code, as can the RNA and protein sequences derived from these genetic instructions.

With this standardized biological manual, molecular structures are consistent in the ways they respond to compounds, set signaling pathways into motion and more. Discovering patterns in these sequences through LLMs can effectively associate yet-untested compounds with certain properties, ranging from targeted drug delivery to immune activation.

Based on these patterns, GenAI can also be used to design novel molecular structures as possible therapies. Scientists can draw inspiration from receptor conformations on viruses or tumor cells to tweak drug compounds and improve their efficacy and safety profiles.

In South Korea, for example, researchers from the Daegu Gyeongbuk Institute of Science & Technology are maximizing such HPC-enabled creative capabilities to explore new candidate proteins for infectious diseases and

neurological disorders. Comprising a cluster of V100-GPU cores, their high-performance supercomputing facility expedited the design of a drug to target Interleukin-1 receptor antagonist, which is a key regulator of immunity and inflammation.

Upon testing in the lab using cellular assays, the team found that their designer anti-inflammatory drug showed strikingly better performance than an approved COVID-19 medication, Anakinra. With the first phase of the development process already complete, trials using *in vivo* pre-clinical models are next on the horizon.

At RIKEN, the AI/HPC pharmaceutical division is also developing a sophisticated platform to span drug discovery until validation, in partnership with several pharmaceutical companies.

“It is not just a single probe or single software,” Matsuoka explained. “The pipeline involves over 50 components, combining software programs, databases and AI algorithms, to generate drug candidates and run simulations to validate the effectiveness of the candidate versus potential dangers.”

**“Drugs are very expensive to develop and the entire process can take 10 to 15 years. One way for costs to go down is by introducing automation and shortening the development cycle.”**

**Professor Satoshi Matsuoka**

Director of the RIKEN Center for Computational Science, Japan





**“[CHROMA and the new innovation center] will catalyze new partnerships between innovators and industry partners, generate new ideas, prototypes and smart technologies for better disease prevention, diagnoses and treatment.”**

**Professor Ivy Ng**  
Former Group CEO of SingHealth, Singapore

## A NEW GENERATION OF DIGITAL DIAGNOSTICS

Just as predicting molecular interactions is no easy feat, evaluating a person's risk for developing disorders is similarly a challenging undertaking.

When assessing heart disease risk, for example, cardiologists must take into account a spectrum of factors, such as age, cholesterol levels and symptoms of chest pain. Co-morbidities like diabetes as well as lifestyle habits like smoking and lack of exercise also contribute to this risk score.

“In general, physicians can roughly predict disease risk, but the accuracy margin is wide,” said Clinical Professor Yeo Khung Keong, CEO and Senior Consultant at the National Heart Centre Singapore, as well as the Academic Chair of the SingHealth Duke-NUS Cardiovascular Sciences Academic Clinical Programme.

At SingHealth, the recently launched AI for the Transformation of Medicine program is poised to bridge this gap, accelerating innovations in smart healthcare through HPC technologies. The Singapore General Hospital campus houses SingHealth's first-ever supercomputer, CHROMA, which is dedicated to processing vast amounts of clinical data and training AI models for biomedical applications.

Jointly developed with the National Supercomputing Centre Singapore, CHROMA is equipped with 1,024 CPU cores and an NVIDIA DGX 320 GB AI accelerator, and is envisioned to facilitate the development of AI models that can predict disease risk and patient trajectories, as well as support health workers in delivering better care to those most in need.

CHROMA is already making waves in the cardiovascular field, as it is being used to train an AI model that can assess a person's risk for a serious cardiac event such as a heart attack. The project, dubbed APOLLO, is a collaboration between the National Heart Centre Singapore at SingHealth; the Agency for Science, Technology and Research; Duke-NUS Medical School; National University Hospital; and Tan Tock Seng Hospital.

“[CHROMA and the new innovation center] will catalyze new partnerships between innovators and industry partners,

generate new ideas, prototypes and smart technologies for better disease prevention, diagnoses and treatment,” said Professor Ivy Ng, who has since stepped down as Group CEO of SingHealth, in a press release.

Once trained, the AI tool will be able to analyze CT scans of the heart's vessels to detect narrowing and plaque build-up, which are critical signs that a person may be at risk for developing cardiovascular disease or experiencing a cardiac event in the future. What makes HPC-enabled AI particularly powerful is the possibility to combine different data types, Yeo noted.

For example, the model could learn to take into account other possible biomarkers including the fatty acid composition of the plaques or data from wearables like the heart rate measured by smart watches, especially contextualized to Asian populations. Moreover, CHROMA on its own can shorten the training phase to just one to two months, compared to the half-year it would typically take.

“What AI brings to the table is reducing the variability between assessing risk scores and increasing the speed of getting the reports,” said Yeo. “We want highly reproducible and consistent tools to raise the accuracy of diagnostics.”

These risk assessments can then help guide clinical decision-making, triaging patients with cardiovascular disease and prioritizing those at high risk for serious cardiac events. The APOLLO team envisions that the integration of such technology in the healthcare workflow can lead to better allocation of hospital resources and the timely delivery of potentially life-saving interventions.

Through AI-powered image analytics, digital pathology solutions are also set to revolutionize cancer detection. Typically, tissue samples from patients are laid on microscopy slides for seasoned pathologists to carefully scrutinize under a microscope. However, tiny cancer cells are not easily recognizable, complicating doctors' efforts to diagnose and assess disease prognosis.

To empower physicians and patients alike, Microsoft and AI company Paige are embarking on a visionary collaboration to develop an image-based GenAI platform that would act as a highly sensitive radar system to spot these malignant cells.

By providing clinical-grade AI and driving the digitalization of modern pathology, the project has the potential to significantly enhance the accuracy and efficiency of clinical oncology work—ultimately enabling precision diagnosis and improving patient outcomes.

## IMPACT AND INTEGRATION

With a vision to build healthier communities, a growing number of countries and institutions are investing in supercomputing resources for biomedical purposes. As HPC-powered healthcare applications pick up speed, it is only a matter of time before these endeavors lead to tangible outcomes for patients.


However, achieving such real-world impact will hinge upon not just technological advancements, but also intentionality and governance over their use. Considering the sensitivity of medical information, Yeo emphasized that regulatory frameworks and practical guidelines must also adapt and evolve alongside these innovations.

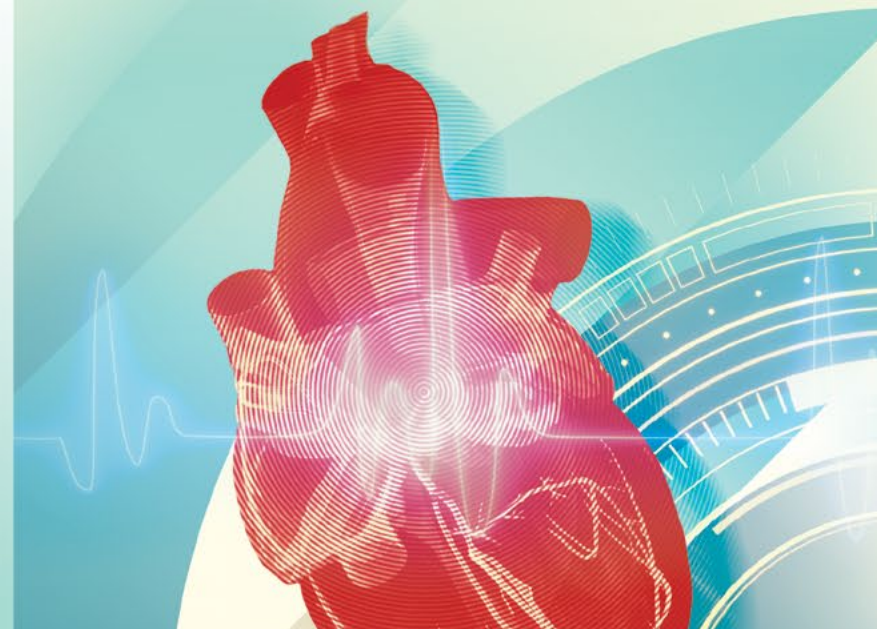
“We can aim for integrating AI in healthcare, but because these technologies would affect lives, there must be safeguards and enough evidence that prove their efficacy and safety,” he said.

Accordingly, research teams are incorporating additional security features and privacy-preserving techniques into their workflows, highlighted Matsuoka. Federated learning is one way to keep local databases separate and inaccessible from other users, while still maximizing the efficiency of the model training process on a global server.

Ensuring ethical use and building trust will become important facets to get physicians and patients on board when it comes to integrating novel technologies in the healthcare system.

When anchored on the values of responsible tech, HPC-enabled AI innovations have the power to transform the future of data-driven and needs-based smart medicine—ranging from the lab-centered beginnings of drug discovery, to the ripples of social impact brought about by enhanced diagnostics and clinical care delivery.

“The biggest thing is to integrate AI technologies into our regular workflows so that we hardly even notice it,” said Yeo. “Whether it's making clinical decisions, monitoring health and alerting patients to warning signs, or predicting outcomes in real-time, supercomputing capabilities would be tremendously important for delivering appropriate medical interventions on an individual level.” 



# AN EYE ON AI

5 Major Breakthroughs Fueled by HPC

Backed by immense computing power, breakthroughs in AI are transforming multiple facets of society, from the way we deliver patient care to how we harness renewable energy resources like solar power.

By **Chia Pei Ling**

Illustrations by Wong Wey Wen / Supercomputing Asia

**In 1982, a young actor David Hasselhoff sits in a self-driving car and gives instructions with just a voice prompt. The car—Knight Industries Two Thousand, or KITT—was arguably one of the first popular television depictions of artificial intelligence (AI), responding to Hasselhoff’s character like a modern-day Alexa. What was once a lofty sci-fi dream has been made possible by advances in AI, backed by the breakneck speeds of high-performance computing (HPC).**



Today, enormous amounts of computing power are used to build generative AI (GenAI) models, which are trained on terabytes of data and have parameters in the billions—with some models reaching the trillion-parameter mark.

At the same time, HPC is also becoming more democratized as cloud computing creates easier access to innovative AI tools and gives enterprises the ability to create their own AI-driven solutions.

Coming full circle, supercomputers are leveling up their own capabilities, powering AI-driven design tools that accelerate the development of increasingly sophisticated semiconductor chips.

## BATTLE OF THE BOTS

Since OpenAI’s ChatGPT tool exploded onto the scene in late 2022, a flurry of large language models (LLMs) has been released as tech giants duke it out to create GenAI capable of accomplishing more complex tasks with greater accuracy.

Across the Pacific, China’s GenAI space is booming. Many Chinese tech firms have built their own proprietary LLMs to give their existing products and services a GenAI-powered refresh, as well as to provide AI-based business solutions on the cloud. As of October 2023, the country’s tech sector has produced at least 130 LLMs—40 percent of the global total.

In March 2023, Baidu, one of China’s largest internet companies, released its own chatbot—Ernie Bot. Since then, Ernie Bot has gained traction, amassing 45 million users since its further release for public use five months later. With a voice prompt, Ernie Bot can create a TV commercial, solve complex geometry problems and even write a martial arts novel filled with twists and turns.

Tapping on its existing userbase, Baidu is using Ernie to redesign and rebuild its products and services, which include Baidu’s search engine, Baidu Maps, and cloud computing service Baidu Cloud. According to a Baidu representative, the Ernie foundation model was built on trillions of data points and hundreds of billions of knowledge points.

In October 2023, Baidu unveiled the latest iteration of its chatbot, Ernie 4.0. At the launch, Baidu CEO Robin Li demonstrated how text- and voice-responsive AI assistants can provide customized search results, navigate a city and add subtitles to videos on a cloud drive.

Beyond products for individual consumers, Baidu has created *Qianfan*, a Model-as-a-Service (MaaS) cloud platform for AI models, targeted at enterprises across diverse sectors including finance, marketing and media.

According to Li, this business model contrasts with other public cloud services that focus on providing computing power and storage, as *Qianfan* also offers businesses access to foundation models built by both Baidu and other third parties. *Qianfan* users can use their proprietary data to fine-tune these pre-installed LLMs, creating tailored solutions for their needs.

Baidu is not the only contender in the LLM arena. The rise of GenAI in China has been dubbed the “war of a hundred models” by Jie Jiang, Vice President of Chinese tech giant Tencent, which released its own model, *Hunyuan*, in September 2023. Developed with businesses in mind, *Hunyuan* is available to Chinese enterprises via Tencent’s cloud platform. The company has also integrated *Hunyuan* into its own products, such as popular mobile messaging app WeChat.

Another player in the space is Alibaba, famed for popular e-commerce platforms like Taobao. In April 2023, Alibaba joined the GenAI frenzy with *Tongyi Qianwen*, updating it to a version 2.0 release in October. Hundreds of millions of Taobao shoppers now have a personal assistant that converses with them to provide tailored product recommendations. Alibaba is also collaborating with New Zealand-based metaverse company Futureverse to train the latter’s text-to-music generation model, JEN-1, on the former’s updated Platform for Artificial Intelligence (PAI).



## THE SUPERPOWERED SEARCH FOR DRUGS

AI is also guiding biotech companies in the quest for the next groundbreaking medical treatment, a traditionally taxing and costly process. Typically, scientists first go on a scavenger hunt for a molecular target in the human body. After setting their sights on a target protein or gene, they may screen up to millions of chemical compounds before landing on a few promising hits. These then need to be optimized in the lab before experimental testing in animal models. These stages of drug discovery—which are completed before starting first-in-human trials—can take up to six years and cost over US\$400 million to get one viable drug candidate.

To accelerate the drug discovery process, AI might take over the heavy lifting for many of these steps. In fact, the world's first fully AI-generated small molecule drug, developed by Hong Kong-headquartered biotech company Insilico Medicine, is currently in Phase 2 clinical trials to evaluate effectiveness after it cleared human safety trials in mid-2023. This path to human trials took less than 30 months at just one-tenth of the typical cost.

To design the drug—aimed at a chronic lung disease called idiopathic pulmonary fibrosis—Insilico used a full suite of AI-driven tools to tackle steps from target discovery to compound generation. In particular, the company created a GenAI drug design engine called Chemistry42,

which churns out never-before-seen molecular structures within days. The fully automated platform is powered by NVIDIA V100 Tensor Core GPUs and can be deployed both in the cloud and on site.

Meanwhile, in Japan, the RIKEN Center for Computational Science (R-CCS) and Fujitsu are developing a next-generation IT drug discovery technology with the help of Asia's fastest supercomputer, Fugaku.

A key aspect of designing optimized drugs is making sure they bind effectively to their target proteins, which makes modeling drug-protein interactions a crucial step in the process. However, proteins are incredibly flexible, toggling between many different conformations and often undergoing significant structural changes when bound to other molecules.

The R-CCS and Fujitsu collaboration addresses the challenge of protein flexibility by combining Fujitsu's deep learning and RIKEN's AI drug discovery simulation technologies. By the end of 2026, the project aims to deliver technology that can analyze drug-protein complexes and predict large-scale structural changes in molecules with high speed and accuracy.

## PETASCALE POWER FOR PHYSICIANS

Besides supporting biomedical research, HPC is also bringing these discoveries from bench to bedside. Singapore's newest petascale supercomputer, Prescience, powers AI models designed to tackle the country's healthcare needs. The fruit of collaboration between the National University Health System (NUHS) and the National Supercomputing Centre (NSCC) Singapore, Prescience's infrastructure is housed at the National University Hospital (NUH) and has been up and running since July 2023. Its on-premise construction obviates the need to de-identify patient data from massive datasets, speeding up model training and enhancing patient data protection.

Packed with multiple NVIDIA DGX A100 compute nodes for the GPU horsepower to handle colossal amounts of data, Prescience is tailored for training LLMs such as RUSSELL—a ChatGPT equivalent for healthcare professionals. Apart from automating administrative tasks like summarizing clinical notes and writing referral letters, RUSSELL also contains NUHS protocols, medical information and rosters to support clinicians in daily tasks.

To help doctors better plan patient treatment and optimize resource allocation, researchers are also

using Prescience to build a patient trajectory prediction model. The model feeds doctors' notes and tests from a patient's medical emergency and first day of inpatient admission to estimate the patient's length of stay. Importantly, the model also provides explainable factors used in its prediction that doctors can easily understand.

Beyond helping clinicians streamline workflows, Prescience is also helping dental patients attain picture-perfect smiles. Through the Smart Monitoring and Intelligent Learning for Enhancing oral health (SMILE AI) project, Singapore's National University Centre for Oral Health (NUCOHS) has been collecting hundreds of dental images to build machine learning models to speed up the routine task of tooth charting and predict the risk of gum disease.

Using X-rays of the upper and lower jaw, NUCOHS's gum disease prediction model aims to stratify patients by their risk of disease. In a push toward preventive healthcare, the model could be implemented on a population level, allowing dentists in the wider community to intervene before disease onset.

"These models are expected to support both clinicians and patients to achieve better outcomes as well as reduce wait times and costs," said Professor Ngiam Kee Yuan, NUHS Group Chief Technology Officer, in an interview with *Supercomputing Asia*.

## CLIMATE CRYSTAL BALLS

Even with such medical advances, healthcare systems in some regions experience added strain from record-breaking heatwaves, which have led to widespread hospitalization and even deaths. One such scorcher that swept across Asia in 2023 saw many countries log temperatures soaring past 40°C, with China's Xinjiang province hitting a searing 52.2°C in July. Such extreme weather events have become more frequent due to climate change, cutting agricultural yields and flooding communities.

To help mitigate damage from such adverse events, AI-driven global climate models provide projections which can aid the design of suitable counter-strategies.

**"These models are expected to support both clinicians and patients to achieve better outcomes, reduce wait times and costs."**

**Professor Ngiam Kee Yuan**  
Group Chief Technology Officer  
of the National University Health System (NUHS), Singapore



That said, the resolution of such global models—which broadly divide Earth into 3D grid cells ranging from 150 to 280 km—often lacks detailed information on regional climates.

To overcome these limitations, the Pawsey Supercomputing Centre (PSC) in Australia is creating high-resolution 3 km grid models for Western Australia (WA), a global biodiversity hotspot. Pawsey's effort is part of the Climate Science Initiative (CSI), a multi-institutional partnership which also includes the WA Department of Water and Environmental Regulation, Murdoch University and the New South Wales Government.

"With finer resolution, we will be able to more accurately predict when and where adverse climate events, such as bushfires and floods, will impact the region. We can also develop better tools to predict the impact of those events," said Mark Stickells, PSC Executive Director, in an interview with *Supercomputing Asia*.

The project is an ambitious one. The team strives to provide comprehensive 75-year climate projections by running simulations from the years 1950 to 2100. With each simulation having two future climate scenarios and two modeling configurations, this entails an immense amount of computing power.

**"With finer resolution, we will be able to more accurately predict when and where adverse climate events, such as bushfires and floods, will impact the region. We can also develop better tools to predict the impact of those events."**

#### Mark Stickells

Executive Director of Pawsey Supercomputing Centre (PSC), Australia

Taking on this gargantuan task is Pawsey's Setonix, the most powerful supercomputer in the Southern Hemisphere, making one-second calculations that would take humans 1.5 billion years to achieve. Setonix is dedicating 40 million core hours and 1.54 petabytes of storage space to CSI—one of the supercomputer's largest allocations.

According to Stickells, findings from CSI will inform government policies for biodiversity conservation and climate-sensitive industries like agriculture. For example, predictions of when seasons will shift can help farming communities predict crop production and manage long-term farming approaches.

## STRIVING FOR SILICON SUCCESS

Even as HPC makes predictions to safeguard our future, it also advances the hardware that powers our present. Semiconductors keep the technological world running, from pocket-sized smartphones to massive supercomputing centers. For example, Japan's Fugaku is powered by 158,976 Fujitsu-designed A64FX semiconductor chips working in tandem. Each system-on-chip contains 48 computing cores with two or four assistant cores, serving as a powerful HPC-tailored processor.

Rapid advances in AI call for more computing power under tight deadlines, and chipmakers are constantly innovating to meet this demand through more advanced silicon chip designs and more efficient manufacturing workflows. Today's semiconductors and supercomputers exist in a symbiotic relationship, with AI stepping in to assist its own makers.

Some of the world's largest chip manufacturers, such as Taiwan Semiconductor Manufacturing Company and Samsung, have leveraged electronic design automation (EDA) to streamline their processes. These Asian chipmakers have partnered with EDA company Synopsys, creators of AI-driven tools to support human engineers in figuring out where to lay out billions of transistors onto tiny silicon pieces. This blueprint is critical as the exact placement of transistors affects a chip's performance.

With chips becoming increasingly complex, engineers often conduct months of manual iterative experiments to land on the best designs for different goals. EDA narrows down design options so engineers focus on the most promising—reducing experimental workload and time taken. Synopsys also provides its EDA software, Synopsys DSO.ai™, on Microsoft's cloud computing platform Azure, allowing companies to

leverage HPC for faster, better results. With HPC and AI support, chipmakers are creating next-generation chips that provide greater speed and consume less power.

However, success with silicon is not infinite. The material transmits light and conducts electricity poorly, making it less-than-ideal for optoelectronic devices like solar cells. At present, the energy conversion efficiency of pure crystalline silicon solar cells has a theoretical limit of 29 percent.

In search of a promising alternative, many scientists are turning to perovskites, a class of crystalline compounds with excellent light absorption properties. By layering perovskites on top of silicon, a perovskite-silicon tandem (PST) solar cell can absorb different wavelengths of light, leading to a higher theoretical efficiency of 43 percent. That said, building a stable and efficient PST solar cell is incredibly challenging as there are approximately  $5^{72}$  possible permutations in a tandem device stack.

In a study published in *Nature Energy*, researchers from South Korea's Chonnam National University led an international collaboration to fabricate a solar cell by stacking two different crystalline structures (or polymorphs) of the perovskite cesium lead iodide ( $\text{CsPbI}_3$ ).  $\text{CsPbI}_3$  has four different polymorphs, two of which are light-absorbing and promising for solar cells. However, the light-absorbing polymorphs can easily convert to non-light-absorbing ones at room temperature, compromising solar cell efficiency.

Through computational simulations by the Roar Supercomputer at Pennsylvania State University, US, the team found that bringing the two light-absorbing polymorphs of  $\text{CsPbI}_3$  together could form a stable atomic interface without distortion. This property allowed the researchers to create a solar cell with a high efficiency of almost 22 percent, which could be stably maintained after 200 hours of storage under ambient conditions.

With such advancements, HPC has evolved from the first 3-megaflop supercomputer in 1964 to the exascale supercomputers we have today. In parallel, the chess- and checkers-playing AI programs of the early 1950s have given way to LLM-powered chatbots. Hand-in-hand, HPC and AI will no doubt continue to make leaps in the decades ahead. ■



In an interview with *Supercomputing Asia*, Professor Torsten Hoefler covers HPC education, regulation and international collaboration.

By **Mitchell Lim**

Illustrations by Wong Wey Wen / *Supercomputing Asia*

# DEFINING THE FUTURE OF AI

A chat with Torsten Hoefler  
of the Swiss National  
Supercomputing Centre



**T**hough not originally designed to function in tandem, high-performance computing (HPC) and artificial intelligence (AI) have coalesced to become a cornerstone of the digital era, reshaping industry processes and pushing scientific exploration to new frontiers.

The number-crunching prowess and scalability of HPC systems are fundamental enablers of modern AI-powered software. Such capabilities are particularly useful when it comes to demanding applications like planning intricate logistics networks or unravelling the mysteries of the cosmos. Meanwhile, AI similarly enables researchers and enterprises to do some clever workload processing—making the most out of their HPC systems.

“With the advent of powerful chips and sophisticated codes, AI has become nearly synonymous with HPC,” said Professor Torsten Hoefer, Director of the Scalable Parallel Computing Laboratory at ETH Zurich.

A master of stringing various HPC components together—from hardware and software to education and cross-border collaborations—Hoefer has spent decades researching and developing parallel-computing systems. These systems enable multiple calculations to be carried out simultaneously, forming the very bedrock of today’s AI capabilities. He is also the newly appointed Chief Architect for Machine Learning at the Swiss National Supercomputing Centre (CSCS),

responsible for shaping the center’s strategy related to advanced AI applications.

Collaboration is central to Hoefer’s mission as a strong AI advocate. He has worked on many projects with various research institutions throughout the Asia-Pacific region, including the National Supercomputing Centre (NSCC) in Singapore, RIKEN in Japan, Tsinghua University in Beijing, and the National Computational Infrastructure in Australia, with research ranging from pioneering deep-learning applications on supercomputers to harnessing AI for climate modeling.

Beyond research, education is also always at the top of Hoefer’s mind. He believes in the early integration of complex concepts like parallel programming and AI processing systems into academic curricula. An emphasis on such education could ensure future generations become not just users, but innovative thinkers in computing technology.

“I’m specifically making an effort to bring these concepts to young students today so that they can better grasp and utilize these technologies in the future,” added Hoefer. “We need to have an education mission—that’s why I’ve chosen to be a professor instead of working in industry roles.”

In his interview with *Supercomputing Asia*, Hoefer discussed his new role at CSCS, the interplay between HPC and AI, as well as his perspectives on the future of the field.

## Q Tell us about your work.

At CSCS, we’re moving from a traditional supercomputing center to one that is more AI-focused, inspired by leading data center providers. One of the main things we plan to do is scale AI workloads for the upcoming “Alps” machine—poised to be one of Europe’s, if not the world’s, largest open science AI-capable supercomputer. This machine will arrive early this year and will run traditional high-performance codes as well as large-scale machine learning for scientific purposes, including language modeling. My role involves assisting CSCS’s senior architect Stefano Schuppli in architecting this system, enabling the training of large language models like LLaMA and foundation models for weather, climate or health applications.

I’m also working with several Asian and European research institutions on the “Earth Virtualization Engines” project. We hope to create a federated network of supercomputers running high-resolution climate simulations. This “digital twin” of Earth aims to project the long-term human impact on the planet, such as carbon dioxide emissions and the distribution of extreme events, which is particularly relevant for regions like Singapore and other Asian countries prone to natural disasters like typhoons.

The project’s scale requires collaboration with many computing centers—and we hope Asian centers will join to run local simulations. A significant aspect of this work is integrating traditional physics-driven simulations, like solving the Navier-Stokes or Eulerian equations for weather and climate prediction, with data-driven deep learning methods. These methods leverage a lot of sensor data we have of the Earth, collected over decades.

In this project, we’re targeting a kilometer-scale resolution—crucial for accurately resolving clouds which are a key component in our climate system.

## Q What is parallel computing?

Parallel computing is both straightforward and fascinating. At its core, it involves using more than one processor to perform a task. Think of it like organizing a group effort among a group of people. Take, for instance, the task of sorting a thousand numbers. This task is challenging for one person but can be made easier by having 100 people sort 10 numbers each. Parallel computing operates on a similar principle, where you coordinate multiple execution units—like our human sorters—to complete a single task.

Essentially, you could say that deep learning is enabled by the availability of massively parallel devices that can train massively parallel models. Today, the workload of an AI system is extremely parallel, allowing it to be distributed across thousands, or even millions, of processing components.

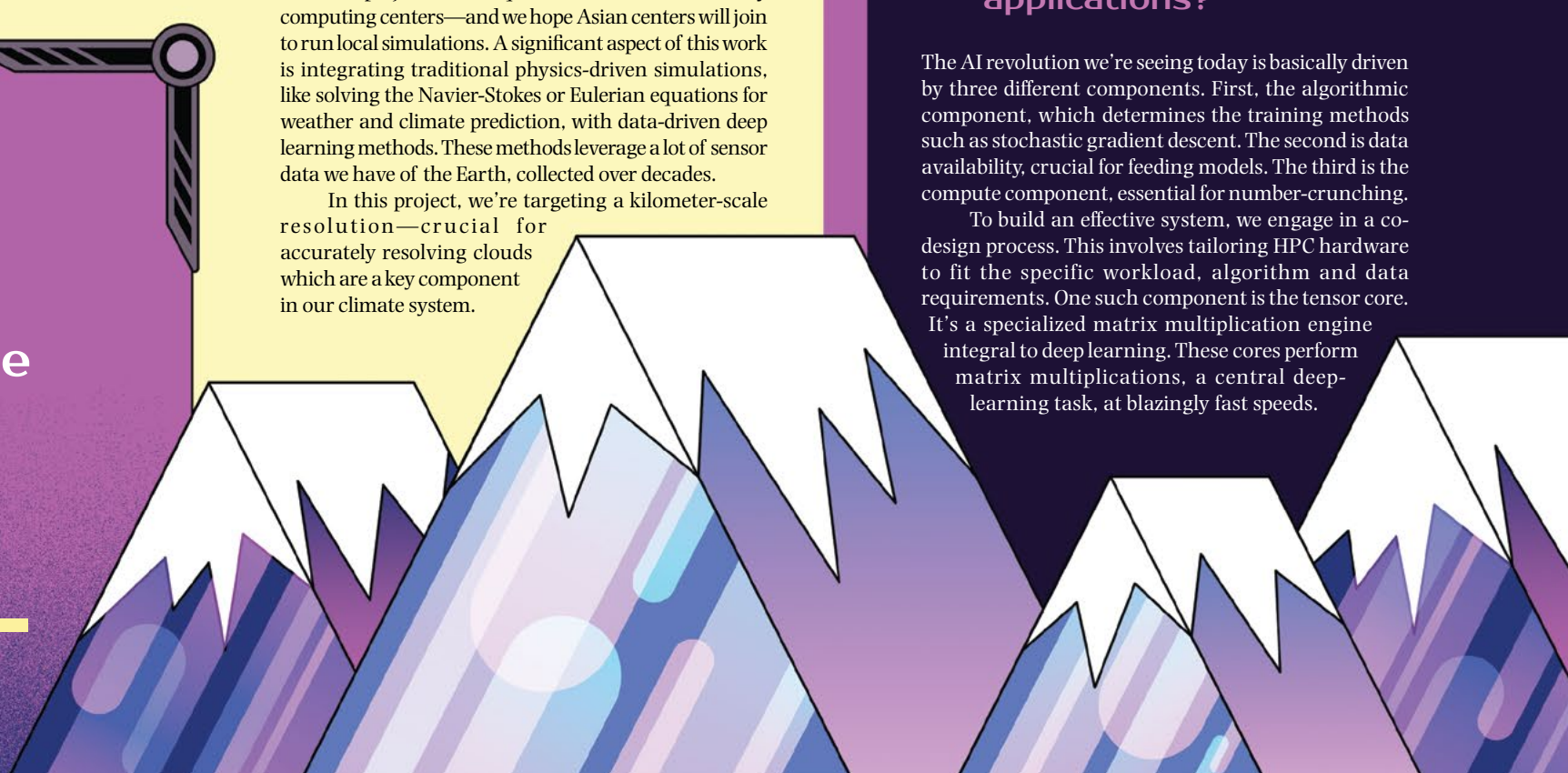
## Q What are some key components for enabling, deploying and advancing AI applications?

The AI revolution we’re seeing today is basically driven by three different components. First, the algorithmic component, which determines the training methods such as stochastic gradient descent. The second is data availability, crucial for feeding models. The third is the compute component, essential for number-crunching.

To build an effective system, we engage in a co-design process. This involves tailoring HPC hardware to fit the specific workload, algorithm and data requirements. One such component is the tensor core. It’s a specialized matrix multiplication engine integral to deep learning. These cores perform matrix multiplications, a central deep-learning task, at blazingly fast speeds.

**“I’m specifically making an effort to bring these concepts to young students today so that they can better grasp and utilize these technologies in the future.”**

**Professor Torsten Hoefer**  
Director of the Scalable Parallel Computing Laboratory at ETH Zurich and Chief Architect for Machine Learning at the Swiss National Supercomputing Centre (CSCS), Switzerland



Another crucial aspect is the use of specialized, small data types. Deep learning aims to emulate the brain, which is essentially a biological circuit. Our brain, this dark and mushy thing in our heads, is teeming with about 86 billion neurons, each with surprisingly low resolution.

Neuroscientists have shown that our brain differentiates around 24 voltage levels, equivalent to just a bit more than 4 bits. Considering that traditional HPC systems operate at 64 bits, that's quite an overkill for AI. Today, most deep-learning systems train with 16 bits and can run with 8 bits—sufficient for AI, though not for scientific computing.

Lastly, we look at sparsity, another trait of biological circuits. In our brains, each neuron isn't connected to every other neuron. This sparse connectivity is mirrored in deep learning through sparse circuits. In NVIDIA hardware, for example, we see 2-to-4 sparsity, meaning out of every four elements, only two are connected. This approach leads to another level of computational speed-up.

Overall, these developments aim to improve computational efficiency—a crucial factor given that companies invest millions, if not billions, of dollars to train deep neural networks.

### Q What are some of the most exciting applications of AI?

One of the most exciting prospects is in the weather and climate sciences. Currently some deep-learning models can predict weather at a cost 1,000 times lower than traditional simulations, with comparable accuracy. While these models are still in the research phase, several centers are moving toward production. I anticipate groundbreaking advancements in forecasting extreme events and long-term climate trends. For example, predicting the probability and intensity of typhoons hitting places like Singapore in the coming decades. This is vital for long-term planning, like deciding where to build along coastlines or whether stronger sea defenses are necessary.

Another exciting area is personalized medicine which tailors medical care based on individual genetic differences. With the advent of deep learning and big data systems, we can analyze treatment data from hospitals worldwide, paving the way for customized, effective healthcare based on each person's genetic makeup.

Finally, most people are familiar with generative AI chatbots like ChatGPT or Bing Chat by now. Such bots

are based on large language models with capabilities that border on basic reasoning. They also show primitive forms of logical reasoning. They're learning concepts like "not cat", a simple form of negation but a step toward more complex logic. It's a glimpse into how these models might evolve to compress knowledge and concepts, like how humans developed mathematics as a simplification of complex ideas. It's a fascinating direction, with potential developments we can only begin to imagine.



### Q What challenges can come up in these areas?

In weather and climate research, the primary challenge is managing the colossal amount of data generated. A single high-resolution, ensemble kilometer-scale climate simulation can produce over an exabyte of data. Handling this data deluge is a significant task and requires innovative strategies for data management and processing.

The shift toward cloud computing has broadened access to supercomputing resources, but this also means handling sensitive data like healthcare records on a much larger scale. Thus, in precision medicine, the main hurdles are security and privacy. There's a need for careful anonymization to ensure that people can contribute their health records without fear of misuse.

Previously, supercomputers processed highly secure data only in secure facilities that can only be accessed by a limited number of individuals. Now, with more people accessing these systems, ensuring data security is vital. My team recently proposed a new algorithm at the Supercomputing Conference 2023 for security in deep-learning systems using homomorphic encryption,

which received both the best student paper and the best reproducibility advancement awards. This is a completely new direction that could contribute to solving security in healthcare computing.

For large language models, the challenge lies in computing efficiency, specifically in terms of communication within parallel computing systems. These models require connecting thousands of accelerators through a fast network, but current networks are too slow for these demanding workloads. To address this, we've helped to initiate the Ultra Ethernet Consortium, to develop a new AI network optimized for large-scale workloads.

These are just some preliminary solutions in these areas—industry and computing centers need to explore these for implementation and further refine them to make them production-ready.

### Q How can HPC help address AI bias and privacy concerns?

Tackling AI bias and privacy involves two main challenges: ensuring data security and maintaining privacy. The move to digital data processing, even in

**“For regulation to be effective, it absolutely must be a global effort. If only one country or a few countries get on board, it just won't work.”**

**Professor Torsten Hoefer**

Director of the Scalable Parallel Computing Laboratory at ETH Zurich and Chief Architect for Machine Learning at the Swiss National Supercomputing Centre (CSCS), Switzerland

sensitive areas like healthcare, raises questions about how secure and private our data is. The challenge is twofold: protecting infrastructure from malicious attacks and ensuring that personal data doesn't inadvertently become part of training datasets for AI models.

With large language models, the concern is that data fed into systems like ChatGPT might be used for further model training. Companies offer secure, private options, but often at a cost. For example, Microsoft's retrieval-augmented generation technique ensures data is used only during the session and not embedded in the model permanently.

Regarding AI biases, they often stem from the data itself, reflecting existing human biases. HPC can aid in “de-biasing” these models by providing the computational power needed. De-biasing is a data-intensive process that requires substantial computing resources to emphasize less represented data aspects. It's mostly on data scientists to identify and rectify biases, a task that requires both computational and ethical considerations.

### Q How crucial is international collaboration when it comes to regulating AI?

International collaboration is absolutely crucial. It's like weapons regulation—if not everyone agrees and abides by the rules, the regulations lose their effectiveness. AI, being a dual-use technology, can be used for beneficial purposes but also has the potential for harm. Technology designed for personalized healthcare, for instance, can be employed in creating biological weapons or harmful chemical compounds.

However, unlike weapons which are predominantly harmful, AI is primarily used for good—enhancing productivity, advancing healthcare, improving climate science and much more. The variety of uses introduces a significant grey area.

Proposals to limit AI capabilities, like those suggested by Elon Musk and others, and the recent US Executive Order requiring registration of large AI models based on compute power, highlight the challenges in this area. This regulation, interestingly defined by computing power, underscores the role of supercomputing in both the potential and regulation of AI.

For regulation to be effective, it absolutely must be a global effort. If only one country or a few countries get on board, it just won't work. International collaboration is probably the most important thing when we talk about effective AI regulation. ■

# The Key To Climate Models

As more countries embrace climate modeling, researchers turn to AI to improve the accuracy and efficiency of their models.

By **Lee Kai Xiang**

Illustrations by Ajun Chuah / *Supercomputing Asia*

The unpredictable nature of the weather has long captured the imagination and fear of our ancestors, immortalized in the legends of gods. From the Grecian Zeus and the Mesoamerican Quetzalcoatl to the Shinto Raijin and Fujin, various deifications have manifested across cultures as humans struggled to impose a sense of order and control to the fickle moods of the winds and clouds.

Modern-day forecasting lifts some of the mysticism around weather with its ability to predict local meteorological conditions, albeit to varying accuracy levels and time horizons. For example, the Meteorological Service Singapore provides fairly accurate projections locally up to a fortnight in advance, while global weather center, AccuWeather, publishes estimates up to three months in advance. The advent of more accurate and extended predictions can help people and governments plan ahead, as well as mitigate property damage and loss of life.

Moreover, the irreversible environmental footprint that human activity has on the planet has led to an increasing global push to understand how the climate changes over much longer timescales. In fact, according to the United Nations Intergovernmental Panel on Climate Change's Sixth Assessment Report, compound weather events—which are combinations of destructive events—will become more frequent as global warming accelerates. The same report highlights that even typical weather events, like maximum daily rainfall and daily temperature extremes, have significantly intensified over the years.

## SEASONAL SUPERCOMPUTING

As governments look to slow the effects of climate change and mitigate its damage, the role of a climate scientist now goes beyond forecasting—they must communicate the full picture and evaluate solutions that help prepare states for any eventuality.

In an interview with *Supercomputing Asia*, Dr M Ravichandran, Secretary of India's Ministry of Earth Sciences, noted that in largely agrarian India, prediction of rainfall patterns has always been paramount to government planning and development of the country. For example, periods of short or extreme precipitation will have to be offset by either building more dams or maintaining water storage and harvesting facilities.

"In the government, we are looking at policies driven by climate information, which will drive the economic sectors, power sectors, agriculture sectors and even negotiation with other governments in the content of policies," Ravichandran shared.

However, weather models are notorious for their appetite for data and computing resources. Scientists must look at multiple chaotic systems that interact with each other, like the atmosphere and the ocean, as well as the influence of space and radiation.

"Because they are so data and compute-intensive, the numerical climate models require full-on massive supercomputers just to simulate snippets of time," noted Dion Harris, head of NVIDIA's data center focusing on high-performance computing, artificial intelligence (AI) and quantum computing.



## THE RIGHT RESOLUTION

Such simulations must be able to account for local weather events and fluctuations to be useful. For instance, a computer model that can only account for spatial weather patterns in a 10 km grid cannot identify the formation of small clouds or local bursts of rain. Additionally, climate scientists must consider the model's time domain: an hourly weather prediction is more useful to the everyday pedestrian than a daily forecast.

In Singapore, the Center for Climate Research Singapore works with the National Supercomputing Centre (NSCC) Singapore to conduct climate studies that address both short- and long-term considerations. The Third National Climate Change Study (V3), which was recently launched, predicts rainfall, temperature, winds and relative humidity at a resolution of 8 km and 2 km up to the year 2100. On top of providing weather information to the public, such data supports the nation's planning for sea levels, water resources, human health, biodiversity and food security.

Given this level of complexity and required resolution, it comes as no surprise that a highly-detailed model combining multiple datasets across decades

is needed to get an accurate picture of the weather. As models become more complex, they also gobble up more computational resources—there are entire supercomputing facilities dedicated to climate research alone. All over Asia, new climate-focused centers are being established as world leaders prepare for turbulent times ahead.

Early last year, Japanese technology giant Fujitsu announced a new supercomputer system provided to the Japan Meteorological Agency for linear rainband forecasting. These slow-moving cumulonimbus clouds bring heavy rains and thunderstorms, which increase the risk of landslides and floods. The new computer features hardware similar to Fugaku, Asia's fastest supercomputer, and will provide more accurate and rapid forecasts.

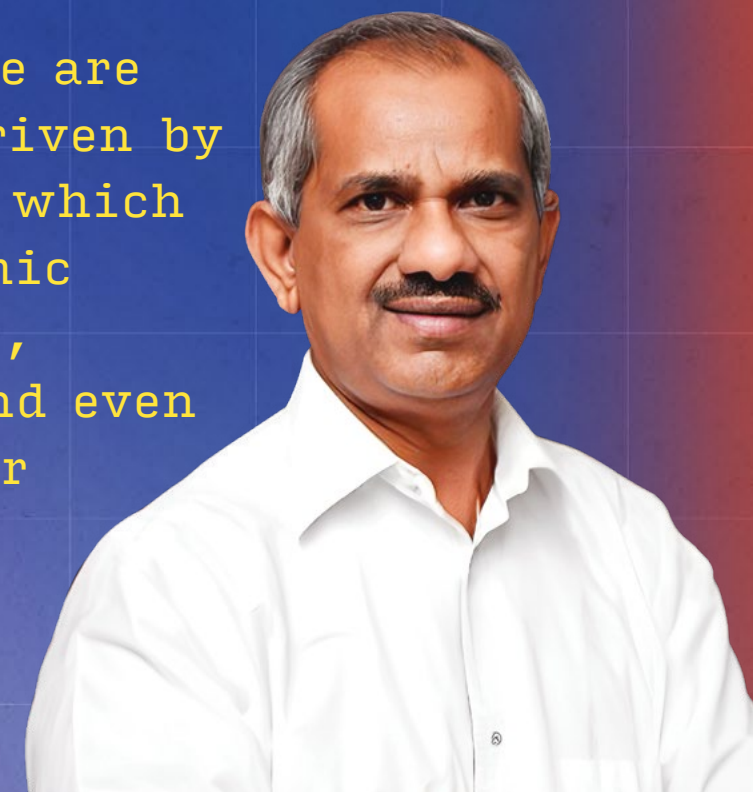
In India, computing solutions company Eviden is working with the Indian Institute of Tropical Meteorology and the National Centre for Medium Range Weather Forecast to deliver two new supercomputers. Higher computing capabilities would allow for better resolution in a virtual model.

"At present, we have a 4 petaFLOPS computer," explained Ravichandran. "We have to achieve 18 petaFLOPS to move from a 12 km resolution to 6 km."

"In the government, we are looking at policies driven by climate information, which will drive the economic sectors, power sectors, agriculture sectors and even negotiation with other governments in the content of policies."

**Dr M Ravichandran**

Secretary of the Ministry of Earth Sciences, India



## OF VIRTUAL EARTHS AND MACHINE LEARNING

In parallel with hardware upgrades for meteorological research, several organizations are looking into techniques used to model the intricacy of long-term climate patterns. Estimates of next month's weather already require a massive effort—imagine predicting years, months or even centuries of climate. Harris underlines the Herculean effort necessary: “You literally would need to be running simulations, all day, every day, for years on end to really get a good view of that.”

To address this issue, researchers and companies are looking into using AI to augment current models. AI is well-known for taking in large amounts of data, spotting patterns and ultimately making fairly efficient and accurate predictions.

At the moment, researchers are confident that AI models will supplement current weather models rather than replace them entirely. Dubbed “digital twins”, current state-of-the-art weather models are computer numerical simulations that construct a virtual diorama of the Earth and her weather patterns.

“To train some of the AI models, you either require a numerical-based simulation to provide most of the data inputs, or you need to simulate some surrogate models.”

In fact, he shared that some scientists are looking into daisy-chaining ensembles of digital twins and AI predictions: running the simulation to provide data for AI, and then using the AI to give economical longitudinal predictions over extended time scales.

However, some might wonder about the reliability of using simulated data to train a machine learning forecast model. Harris provided assurance on the stringent checks and balances put in place, with repeated comparisons performed on both the AI predictions and numerical simulations against real-world events, which are then used to further calibrate such models.

NVIDIA hosts their own data-driven weather model, dubbed the **Fourier Forecasting Neural Network** (FourCastNet). Accelerated by graphics processing, predicting the weather a week in advance only requires a fraction of a second on a single NVIDIA graphics processing unit (GPU).

Another revolutionary model is the Pangu-Weather model developed by Huawei Cloud. Published in *Nature* in 2023, Pangu-Weather breaks new ground as the first AI-backed model to outperform traditional numerical methods.

The model has been extensively tested on various key events, with remarkable success in modeling the sweltering 40°C UK summer in 2022, as well as tracking the path of Storm Eunice in 2022 and Typhoon Doksuri in 2023.

In late 2023, Google released Graphcast, the newest contender in the field. Built upon four decades of data, GraphCast outperforms conventional numerical models, demonstrating precise tracking of Hurricane Lee as well as various extreme thermal events in the same year.

Although such AI-supported forecast models are built upon slightly different architectures, they undeniably reinforce the superior edge AI provides in the arena of weather prediction. All three models are open-source, with charts available to the public on the European Centre for Medium-Range Weather Forecast's website.

## AN OPEN FORECASTING REVOLUTION

Highlighted by a recent paper in *Nature*, active collaborations and data sharing will be the fulcrum for incorporation of larger climate models. Experts are excited to see the rapid proliferation of AI among national weather and climate research centers.

With the integration of AI into climate models, a few GPUs can now run a weather model to a degree of accuracy that previously required a supercomputer to achieve. Current state-of-the-art models can also be run on a bulked-up personal computer, making weather forecasting more accessible than any other point in history.

For governments, having hardware-efficient models opens the possibility of having smaller regional centers for forecasting. Without the huge investment of a new supercomputing center, these centers can be set-up at lower costs, yet perform fairly accurate predictions, bringing better foreknowledge of the weather to the public. Taiwan is currently engaging with NVIDIA to gain a better understanding of the regional consequences of weather events.

At the same time, with climate models going open-source, researchers in the public domain, academia and industry can now work hand-in-hand to develop the next generation of climate models.

Harris highlights that NVIDIA has been working closely with the development community to ensure that its software runs quickly and efficiently. “We’re engaging with the broader community to help optimize and make sure that the models can be easily adopted and fine-tuned to meet reasonable use cases,” Harris said.

With the advent of faster and more accurate models, as well as upgrades to existing computing facilities for climate science, governments can now have a better understanding of the consequences of long-term countermeasures in fighting climate change. The new accessibility of climate science and modeling forecasts clear skies ahead for meteorology. ■



# NVIDIA ANNOUNCES A SLEW OF PRODUCTS AND PARTNERSHIPS IN TAIWAN

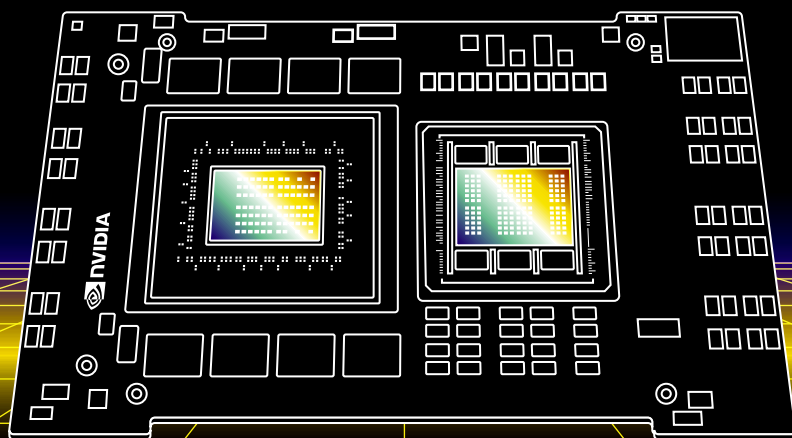
NVIDIA announced plans to work with Taiwanese tech suppliers on artificial intelligence (AI) and automotive electronics advancements. In particular, the chip giant has plans to collaborate with mobile chip developer MediaTek on automotive solutions.

The collaboration will combine MediaTek's system-on-a-chip and NVIDIA's graphics processing unit (GPU) and AI technologies to offer AI, connectivity and computing capabilities to next-generation smart vehicles.

At Computex Taipei 2023, NVIDIA Founder and CEO, Jensen Huang, also announced the GH200 Grace Hopper Superchip.

The superchip is a combined accelerator and processor based on the NVIDIA's Grace CPU and Hopper H100 GPU architectures. It is currently available to Google Cloud, Meta and Microsoft and is expected to appear in systems launched in Q2 of 2024.

"To meet surging demand for generative AI, data centers require accelerated computing platforms with specialized needs," said Huang. "The new GH200 Grace Hopper Superchip platform delivers this with exceptional memory technology and bandwidth to improve throughput, the ability to connect GPUs to aggregate performance without compromise, and a server design that can be easily deployed across the entire data center."



## MALAYSIA ADAPTS TO ATTRACT NEW DATA CENTERS

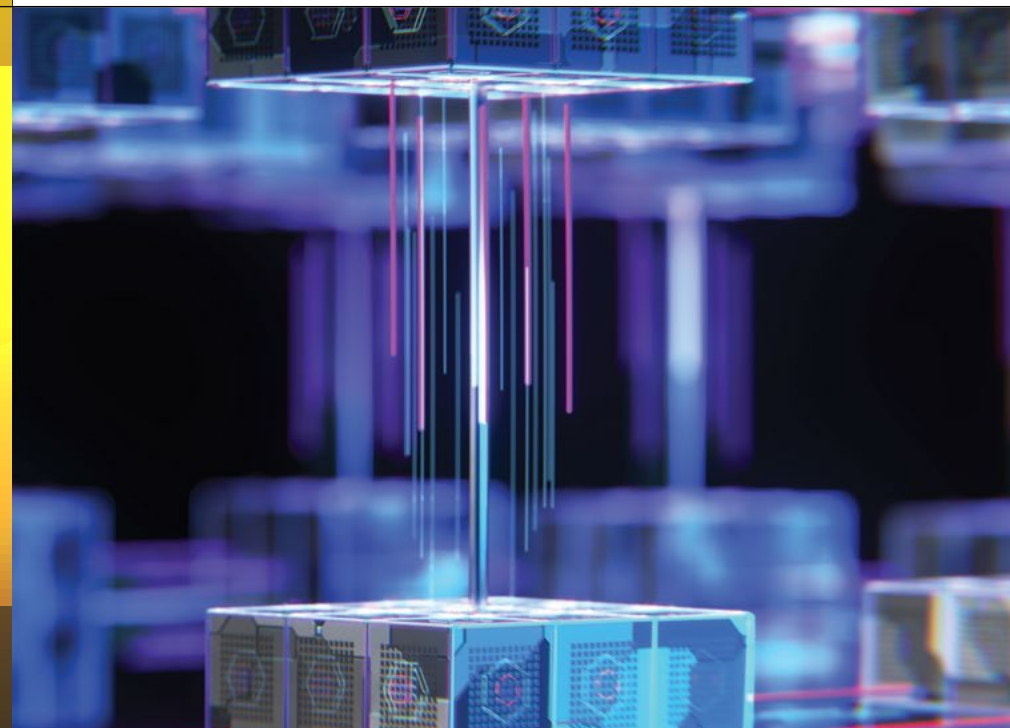
As demand for data centers across Southeast Asia grows with the region's economy, countries like Malaysia work to create an ideal ecosystem for the facilities.

Just a 30-minute drive from Singapore—the region's current data hub—Iskandar Puteri in Malaysia's southern Johor state is set to become a new hot spot for data centers. Malaysia boasts affordable land, electricity and new tax breaks for digital infrastructure investments.

A China-headquartered data center developer, GDS Holdings, opened the doors to its 69.5-megawatt facility in Iskandar Puteri in September 2023—the company's first facility outside China. US data center specialists, Equinix, are currently building a US\$40 million data center in the area as well.

To facilitate such investment, Malaysia has launched two key programs—the Digital Ecosystem Acceleration scheme that waives taxes on qualifying investments as well as the New Industrial Master Plan 2030 which promotes digitization.

"Our target for Malaysia's data center industry is to achieve a revenue of close to US\$800 million by 2025," said Tengku Zafrul Abdul Aziz, Minister of Investment, Trade and Industry.



## HONG KONG SAR SETS ASIDE HK\$50M SUM FOR HPC AND WEB3 DEVELOPMENT

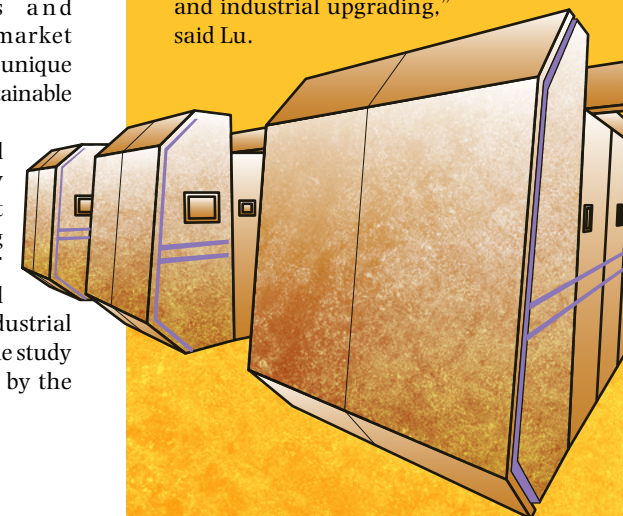
In the 2023–2024 budget, Hong Kong Special Administrative Region (SAR) announced plans to dedicate HK\$50 million (US\$6.39 million) to developing the city's third generation internet (Web3) ecosystem and supercomputing capabilities.

Finance chief Paul Chan Mo-po noted Web3's potential and emphasized the importance of developing virtual assets. Chan plans to create and lead a task force on virtual asset development.

The task force will consist of a diverse set of stakeholders—

from policy bureaus and financial regulators to market participants—each offering unique recommendations on the sustainable development of the sector.

Chan also announced plans to conduct a feasibility study on the development of an AI Supercomputing Center. Proponents of the center believe it will strengthen Hong Kong's industrial and research capabilities. The study is planned to be completed by the end of 2024.



## CHINA UNVEILS DOMESTIC SUPERCOMPUTER TIANHE XINYI

On December 6, 2023, the National Supercomputing Center (NSC) in Guangzhou announced that a new supercomputer, Tianhe Xinyi, has commenced operation. According to NSC, the supercomputer is completely home-developed and operates on domestically designed chips and computing architecture.

While the center did not disclose specific performance parameters, NSC center director Professor Lu Yutong said that the new computer has outperformed Tianhe-2, one of China's fastest supercomputers that was a world leader from 2013 to 2015.

China has two existing exascale supercomputers—Sunway TaihuLight and Tianhe-3. But they have not been submitted for ranking on the Top500 list. Similarly, when announcing the Tianhe Xinyi, the NSC did not highlight its speed and instead chose to emphasize the use of local chips and potential applications.

"It will provide users in the Guangdong-Hong Kong-Macao Greater Bay Area a strong platform and high-end computing power to achieve technological breakthroughs and industrial upgrading," said Lu.

Super Snapshot

# CHARTING A RESEARCH ROADMAP

Quek Gim Pew, Chairman of the National Supercomputing Centre (NSCC) Singapore and winner of the President’s Science and Technology Medal 2023, has built an illustrious career that has supported Singapore’s scientific and technological progress. Find out how his influence and leadership have shaped Singapore’s research ecosystem in areas like artificial intelligence (AI), defense technology, high-performance computing (HPC) and quantum engineering.



2004   2016	<p>Appointed CEO of DSO National Laboratories (DSO) after two decades of service</p> <p><i>At DSO, Quek’s leadership fostered an environment where the operational needs of the Singapore Armed Forces (SAF) were addressed with innovative solutions.</i></p>
2015	<p>Joined the NSCC steering committee</p>
2016   2021	<p>Appointed Chief Defence Scientist of the Singapore Ministry of Defence (MINDEF)</p> <p><i>Here, Quek designed the masterplan of defense research and development (R&amp;D), solidified local and international partnerships, and drove the development of STEM talent in defense.</i></p> <p><i>“MINDEF leadership showed a healthy appetite for high-risk, high-payoff R&amp;D—but demanded that payoffs must support the operational needs of the SAF. I suppose my background in this environment gave me a strong sense of purpose and confidence in our ability to harness technology to solve seemingly intractable problems.”</i></p>
2016	<p>Appointed Governing Board Chair of the Center for Quantum Technology (CQT)</p> <p><i>Quek’s foresight on the importance of quantum technology helped increase the country’s interest in application-based research that harnesses the maturing technology for economic and national goals.</i></p>
2017	<p>Appointed Scientific Committee Chairman of AISG</p> <p><i>“AI is not new to the local research community, but the development of generative AI has certainly injected a lot of buzz and excitement; so much so that our HPC infrastructure was not able to cope with the exponential increase in demand!”</i></p>
2021	<p>Appointed Deputy Chairman of the Office for Space Technology and Industry (OSTIn)</p> <p><i>Similarly, Quek’s vision has helped secure access to space technologies, maximize economic value, and increase strategic relevance in the field.</i></p> <p>Appointed Co-Chair of the National Quantum Steering Committee (NQSC)</p> <p><i>“We should take a close look at quantum. How can we prepare ourselves? It is fortunate that we have a very strong quantum research community here that allows us to understand the technology, distinguish reality from hype, and develop a robust roadmap ahead.”</i></p>
2023	<p>Appointed Chairman of NSCC</p> <p><i>“The evolution of HPC organization in Singapore reflects how HPC has impacted research, development and industry.”</i></p>

# SUPERCHARGE YOUR CAREER WITH HIGH PERFORMANCE COMPUTING (HPC) SKILL SETS

## CERTIFICATE OF COMPETENCY (COC) IN INTRODUCTION TO HPC

Embark on your HPC Learning journey today, facilitated by the collaboration between the **National Supercomputing Centre (NSCC) Singapore** and **ITE College West**.

Gain fundamental HPC knowledge and kickstart your learning journey today!

Use your SkillsFuture credits:

Scan to register NOW!

Singapore Citizens & Permanent Residents  
**\$62.13**

Singapore Citizens aged 40 & above  
**\$24.13**

Non Citizens (Full Fee)  
**\$207.10**



For more info:  
6590 2628  
college\_west@ite.edu.sg



# REDEFINING THE FUTURE OF HPC

## NSCC AND ALTAIR COLLABORATE TO CONNECT HPC COMMUNITIES AND OPTIMIZE RESOURCES

As long-time Altair users, the National Supercomputing Centre (NSCC) Singapore chose to work with Altair to test and develop a solution that would let users run workloads using a global pool of resources with Altair® Liquid Scheduling™. Liquid Scheduling extends the Altair HPC stack already deployed at NSCC, taking the Altair® PBS Professional® workload manager to another level of scalability and performance.

Liquid Scheduling is a powerful, flexible HPC feature that meets the demands of the latest distributed workflows. It ensures that workloads run in the most efficient manner by connecting multiple clusters and sites, eliminating silos, and providing global visibility into resource utilization.

**Read the full customer story - [altair.com/nscc-liquid-scheduling](https://altair.com/nscc-liquid-scheduling)**



© Altair Engineering, Inc. All Rights Reserved. / [altair.com](https://altair.com) / Nasdaq: ALTR

[in](#) [f](#) [t](#) [@](#) #ONLYFORWARD