ICPP 2026

The 55th International Conference on Parallel Processing

28 September – 1 October 2026
NTU@one-north
11 Slim Barracks Rise, Singapore 138664

Welcome to ICPP2026

Welcome to the 55th International Conference on Parallel Processing (ICPP 2026) in Singapore!

For more than five decades, ICPP has served as a premier forum for researchers, practitioners, and industry leaders to exchange ideas and advance the state of the art in parallel and distributed computing. We are delighted to bring the ICPP community to Singapore from September 28 to October 1, 2026, and to present a rich technical program reflecting both the foundations of our field and its rapidly expanding role in artificial intelligence and emerging computing platforms.

The ICPP 2026 program features 117 full papers, 8 posters, 4 demonstrations, 13 workshop papers, and 2 tutorials. Together, these contributions represent a broad spectrum of advances in parallel processing, from fundamental algorithms and architectures to high-performance systems, AI infrastructure, and emerging computing technologies.

We are particularly pleased to feature four distinguished keynote speakers whose talks capture important directions shaping the future of parallel computing:

Xian-He Sun, Illinois Institute of Technology — *From Scalable Computing to the Scaling Law: A Reexamination of Parallel Processing*

Xiaowen Chu, The Hong Kong University of Science and Technology (Guangzhou) — *Efficient LLM Serving via Compression-Centric Optimizations*

Torsten Hoefler, ETH Zurich — *Ultra Ethernet for Next-Generation AI and HPC Workloads*

Li-Shiuan Peh, National University of Singapore — *Parallel Computing on Ultra-Low-Power Chips*

These keynotes highlight the remarkable breadth of the field today—from scalability and data-centric computing, to efficient large-language-model serving, next-generation AI and HPC interconnects, and energy-efficient computing at the edge.

A conference of this scale is only possible through the efforts of a large and dedicated community. We sincerely thank the track chairs, program committee members, reviewers, and session chairs for their tremendous commitment to the technical program and for the time and care they devoted to the review and selection process. We are equally grateful to the workshop, tutorial, poster, demo, publicity, web, finance, proceedings, local organization, and other chairs and volunteers whose contributions made ICPP 2026 possible.

We also thank all the authors who chose ICPP as the venue to share their work. Whether or not a submission ultimately appeared in the program, the quality and diversity of the submissions demonstrate the vitality of the parallel processing community. We especially congratulate the authors whose work is presented at ICPP 2026.

Beyond the technical program, we hope that ICPP 2026 provides many opportunities for discussion, new collaborations, and reconnecting with colleagues from around the world. Singapore—with its vibrant research ecosystem, strong connections between academia and industry, and unique position at the crossroads of Asia and the world—provides an exciting setting for these interactions.

We hope you enjoy the technical program, the conversations, and your time in Singapore.

General Co-Chairs

Wentong Cai, Nanyang Technological University, Singapore
Bingsheng He, National University of Singapore, Singapore
Tamar Eilam, IBM T.J. Watson Research Center, USA

Program Co-Chairs

Xueyan Tang, Nanyang Technological University, Singapore
Amelie Chi Zhou, Hong Kong Baptist University, China
Suren Byna, The Ohio State University, USA

Organizing Committee

General Co-Chairs

Wentong Cai
Nanyang Technological University, Singapore

Bingsheng He
National University of Singapore, Singapore

Tamar Eilam
IBM T.J. Watson Research Center, USA

Program Co-Chairs

Xueyan Tang
Nanyang Technological University, Singapore

Amelie Chi Zhou
Hong Kong Baptist University, China

Suren Byna
The Ohio State University, USA

Track Chairs

Algorithms	**Olivier Beaumont**, INRIA, France **Giulia Guidi**, Cornell University, USA
Applications	**Radu Prodan**, University of Innsbruck, Austria **Nishant Saurabh**, University of Utrecht, Netherlands
Architecture	**Ivy Peng**, KTH Royal Institute of Technology, Sweden **Antonino Tumeo**, Pacific Northwest National Laboratory, USA
Multidisciplinary	**Ariful Azad**, Texas A&M University, USA **Helen Xu**, Georgia Tech, USA
Performance	**Jidong Zhai**, Tsinghua University, China **Shadi Ibrahim**, INRIA, France
Software	**Rajeev Thakur**, Argonne National Laboratory, USA **Tanzima Islam**, Texas State University, USA
Quantum Computing	**Qiang Guan**, Kent State University, USA **Bo Fang**, University of Texas at Arlington, USA

Workshop Chairs

Quan Chen, Shanghai Jiao Tong University, China
Jay Lofstead, Sandia National Laboratories, USA

Tutorial Chairs

Chen Wang, Nanyang Technological University, Singapore
Cheng Chen, ByteDance, Singapore

Poster Chairs

Qinbin Li, Huazhong University of Science and Technology, China
Matthieu Dorier, Argonne National Laboratory, USA

Demo Chairs

Yao Chen, Huazhong University of Science and Technology, China
Jean Luca Bez, Lawrence Berkeley National Laboratory, USA

Finance Chair

Jiong He, National University of Singapore, Singapore

Sponsor Chair

Tao Luo, A*STAR Institute of Advanced Intelligence and Computing, Singapore

Proceedings Chair

Nguyen Binh Duong Ta, Singapore Management University, Singapore

Publicity Chairs

Rui Tan, Nanyang Technological University, Singapore
Jidong Zhai, Tsinghua University, China
Ayesha Afzal, Erlangen National High Performance Computing Center (NHR@FAU), Germany

Web Chairs

Weihao Cui, Shanghai Jiao Tong University, China
Junyi Hou, National University of Singapore, Singapore

Keynote Speakers

All keynotes in AUD302, Level 3

KEYNOTE 1 TUESDAY 29 SEPTEMBER 9:00

From Scalable Computing to the Scaling Law: A Reexamination of Parallel Processing

Xian-He Sun, Illinois Institute of Technology, Chicago, USA

AI's recent success is driven by scaling laws, where increasing data and compute power yields predictably better accuracy. This trend demands computing systems that are infinitely scalable. This talk reevaluates parallel processing through both scale-out and scale-up lenses to address these AI-driven challenges. After reviewing foundational breakthroughs in the field, we present DataflowV (Dataflow under the von Neumann machine), a data-centric methodology that maximizes parallelism and concurrency from different aspects. We illustrate the DataflowV concept using two NSF-funded cyberinfrastructure systems: Hermes, for large-scale data handling, and ChronoLog, for application-specific AI-integrated optimizations. We conclude by introducing IOWrap, a data-management system tailored for the unique demands of modern AI agents.

KEYNOTE 2 TUESDAY 29 SEPTEMBER 13:40

Ultra Ethernet for Next-Generation AI and HPC Workloads

Torsten Hoefler, ETH Zurich, Switzerland

The Ultra Ethernet Consortium set out to redefine Ethernet-based interconnects for AI and high-performance computing (HPC), culminating in the recent release of its first specification (version 1.0). We will analyze HPC and AI workloads with respect to their networking requirements. We will continue by highlight key innovations that distinguish Ultra Ethernet from existing solutions, ranging from lossy operation—both with and without trimming—to fully hardware-offloaded rendezvous protocols. We will explore the architectural advancements and technical highlights that enhance efficiency, scalability, and performance, positioning Ultra Ethernet as a transformative force in next-generation computing. We will also discuss current experiences with data from real-world large-scale deployment running jobs on up to 75000 GPUs stably through many switch and link failures in the network.

KEYNOTE 3 WEDNESDAY 30 SEPTEMBER 9:00

Efficient LLM Serving via Compression-Centric Optimizations

Xiaowen Chu, The Hong Kong University of Science and Technology (Guangzhou), China

Large language model (LLM) serving is fundamentally constrained by GPU memory capacity and bandwidth. While model compression can alleviate these constraints, practical system-level speedups are frequently hindered by the computational overhead of decoding and data movement. This talk explores a compression-centric paradigm for efficient LLM serving, highlighting two synergistic systems: SpInfer and ZipServ. We demonstrate how hardware-aware format co-design bridges the gap between theoretical compression and practical acceleration. First, SpInfer optimizes sparsified LLM serving by redesigning sparse data structures to minimize indexing overhead, unlocking the efficiency of unstructured sparsity on GPUs. Second, ZipServ extends this philosophy to bit-exact inference via lossless compression, enabling on-the-fly decoding that strictly aligns with Tensor Core execution to eliminate redundant memory traffic. Ultimately, this talk illustrates that by co-designing compression formats and execution pipelines with GPU architectures, compression transitions from a mere storage-saving technique into a core systems-level optimization principle that simultaneously minimizes memory footprint and accelerates LLM inference.

KEYNOTE 4 WEDNESDAY 30 SEPTEMBER 13:40

Parallel Computing on Ultra-Low-Power Chips

Li-Shiuan Peh, National University of Singapore

This talk will discuss how parallel computing is making inroads into ultra-low-power chips. I will chronicle how the highly parallel architectures of supercomputers and servers that enabled scalable HPC have made their way into AI accelerators in data centers in recent years, and are moving further into edge devices to enable power-efficient chips that deliver high performance at mW. Next, I will walk through recent research into ultra-low-power AI accelerators and several of our chip prototypes targeting battery-powered edge devices. The talk will round off with a glimpse into how these chips can enable next-generation AI-powered wearables.

Program at a Glance

Mon 28 Sep WORKSHOPS + TUTORIALS

09:00–12:30	SPIN-NVSC — State of Practice in Deploying Supercomputers with NVIDIA Superchips	LT701
	SUSTAIN-HPC — Sustainable Computing for High-Performance and Distributed Systems	LT702
	BID — Benchmarking in the Data Center	SR703
	Tutorial — Accelerating AI and HPC Workflows with AMD ROCm AI and Modern GPU Clusters	SR706
12:30–13:30	Lunch	Level 7
13:30–17:00	SPIN-NVSC — State of Practice in Deploying Supercomputers with NVIDIA Superchips	LT701
	Agentic AI in Real-World Systems: Infrastructure, Algorithms, and Deployment	LT702
	DC4AI — Data Compression for AI and Big Data Applications	SR703
	Tutorial — Coyote V2: An Open-Source FPGA Shell for Modern Distributed Data Processing	SR706

Tue 29 Sep MAIN CONFERENCE DAY 1

08:45	Opening			AUD302
09:00	**Keynote 1** — Xian-He Sun: From Scalable Computing to the Scaling Law: A Reexamination of Parallel Processing			AUD302
10:30	**1A** MoE Systems	AUD302	**1B** Disaggregated Memory	LT301
	1C Scientific Computing I	LT701	**1D** Compiler Optimization	LT702
13:40	**Keynote 2** — Torsten Hoefler: Ultra Ethernet for Next-Generation AI and HPC Workloads			AUD302
15:10	**2A** LLM Inference and Serving I	AUD302	**2B** Processing-in-Memory	LT301
	2C Numerical Computing	LT701	**2D** Graph Algorithms	LT702
17:00	Posters + Live Demonstrations			Function Hall (Level 3)
18:00	Reception			Function Hall (Level 3)

Wed 30 Sep MAIN CONFERENCE DAY 2

09:00	**Keynote 3** — Xiaowen Chu: Efficient LLM Serving via Compression-Centric Optimizations			AUD302
10:30	**3A** LLM Inference and Serving II	AUD302	**3B** Key-Value Stores	LT301
	3C Hardware Acceleration	LT701	**3D** Graph Learning	LT702
13:40	**Keynote 4** — Li-Shiuan Peh: Parallel Computing on Ultra-Low-Power Chips			AUD302
15:10	**4A** LLM Inference and Serving III	AUD302	**4B** Storage Systems	LT301
	4C Sparse Linear Algebra I	LT701	**4D** Performance Analysis and Modeling I	LT702
17:15	Shuttle from NTU@one-north to Banquet			
18:30–21:00	Conference Banquet — The Ballroom @ Mount Faber Peak			

Thu 1 Oct MAIN CONFERENCE DAY 3

09:00	**5A** Memory Management	AUD302	**5B** AI Acceleration	LT301
	5C Serverless Computing	LT701	**5D** Parallel Data Processing	LT702
10:45	**6A** AI Caching	AUD302	**6B** Learning-Based Scheduling	LT301
	6C Network Algorithms	LT701	**6D** Performance Analysis and Modeling II	LT702
13:00	**7A** Sparse Linear Algebra II	AUD302	**7B** Data-Center Networks	LT301
	7C Distributed Learning	LT701	**7D** Parallel Training	LT702
15:10	**8A** GEMM Optimization	AUD302	**8B** Cloud Scheduling	LT301
	8C Federated Learning	LT701	**8D** Scientific Computing II	LT702
16:50	Conference closes			

Tuesday, 29 September Morning

08:45–09:00 Opening • AUD302

09:00–10:00 **Keynote 1 — Xian-He Sun**, Illinois Institute of Technology
From Scalable Computing to the Scaling Law: A Reexamination of Parallel Processing • AUD302

10:00–10:30 Tea Break Function Hall (Level 3)

1A — MoE Systems

10:30–12:10 • AUD302

1. MigMoE: Task-Aware Expert Migration for Faster and More Balanced Expert-Parallel MoE Inference
2. CARE-MoE: Correlation-Aware Expert Placement and Semantic Equivalence Routing for MoE LLM Inference on Edge Devices
3. BR-MoE: A Memory-Budget-Aware Co-Design Framework for Efficient MoE Inference on Resource-Constrained GPUs
4. Fine-grained Computation-Communication Overlap via Tile-level Signaling and Scheduling for Mixture-of-Experts

1B — Disaggregated Memory

10:30–12:10 • LT301

1. Aethon: Performance-aware Memory Offloading for Co-running Applications in Public Clouds
2. Duet: Orchestrating Full-Duplex-Aware Proactive Demotion in CXL Tiered Memory
3. MEDO: Adaptive Multi-Stream Data Offloading for Efficient Management of Disaggregated Memory Pool
4. Breaking the Migration Barrier: Hardware-Software Co-Designed Translation Coherence for Tiered Memory

1C — Scientific Computing I

10:30–12:10 • LT701

1. Large-Scale Quantum Circuit Simulation on HPC Cluster via Cache Blocking, Boosting, and Gate Fusion Optimization
2. ZSWP: Rollback-Free Spherical ε-Join for Scalable Astronomical Cross-Matching on Heterogeneous Supercomputer Architectures
3. High-Performance Virtual Screening Based on Parallel Molecular Dynamics Simulations with Early Termination
4. AtomFlow: Accelerating GNN-Based Molecular Dynamics with ab initio Accuracy on Fugaku Supercomputer

1D — Compiler Optimization

10:30–12:10 • LT702

1. In-Copy Fusion: Runtime Argument Fusion for Efficient OpenMP GPU Offloading
2. Adaptive Operator Fusion with Subproblem Decomposition for Enhanced Memory Efficiency
3. Thinking Globally, Acting Locally: Merge Sort's Auto-Vectorization on Shuffle-Heavy ISAs
4. A SYCLic Investigation of SYCL Ecosystems: UniSYCL for Heterogeneous Scaling

12:10–13:40 Lunch Function Hall (Level 3)

Tuesday, 29 September Afternoon

13:40–14:40 **Keynote 2 — Torsten Hoefler**, ETH Zurich
Ultra Ethernet for Next-Generation AI and HPC Workloads • AUD302

14:40–15:10 Tea Break Function Hall (Level 3)

2A — LLM Inference and Serving I

15:10–16:50 • AUD302

1. RPSC: Robust LLM Scheduling by Tolerating Prediction Inaccuracy and Mitigating Tail Latency
2. CrossServe: Cross-Layer Scheduling for SLO Optimization in Multi-Tenant LLM Serving
3. WAQ-LLM: Optimizing Multi-Instance LLM Deployment via Workload-Aware Queueing Model
4. Heterogeneous SLO Guaranteed Multi-Resource-Aware Batching in LLM Serving

2B — Processing-in-Memory

15:10–16:50 • LT301

1. Cuckoo-GPU: Accelerating Cuckoo Filters on Modern GPUs
2. Query Density-Driven Partitioning for Spatiotemporal Load Balancing on Processing-in-Memory Systems
3. PIM-Pai: Boosting the Performance of Index Structures with Minimal Effort on Processing-in-Memory Architecture
4. PIM-Forge: Toward Autonomous PIM Kernel Synthesis through Constraint-Guided Hardware Exploration

2C — Numerical Computing

15:10–16:50 • LT701

1. VILIB++: High Performance C++ Library for Auto-Vectorizing Interval Arithmetic
2. From 2^N to N^2: Tree-Free Scalable Sparse Symmetric Tucker Decomposition
3. High-Performance Star-M SVD for Big Data Compression
4. Elasticity in Parallel Sparse Triangular Solve

2D — Graph Algorithms

15:10–16:50 • LT702

1. Sparsity-aware Fine-grained Parallelization for Set Intersection Computations in Graph Datasets
2. Contraction Hierarchies for Parallel Sequence-to-Graph Alignment
3. SQUASH: Distributed Square Estimation with Provable Error Bounds for Dynamic Graphs
4. DyGMIS: Fast and Scalable Fully Dynamic MIS Maintenance on Large Evolving Graphs

17:00–18:00 **Posters + Live Demonstrations**
Function Hall (Level 3)

18:00–20:00 **Reception**
Function Hall (Level 3)

Wednesday, 30 September Morning

09:00–10:00 **Keynote 3 — Xiaowen Chu**, The Hong Kong University of Science and Technology (Guangzhou)
Efficient LLM Serving via Compression-Centric Optimizations • AUD302

10:00–10:30 Tea Break Function Hall (Level 3)

3A — LLM Inference and Serving II

10:30–12:10 • AUD302

1. ReliefServe: Relieving GPU Pressure in Multi-Model Serving via Selective CPU Escape
2. AsymFlow: Enabling Long-Context LLM Serving via CPU-GPU Prefill-Decode Disaggregation
3. IHS-LM: Intra-batch Hybrid Scheduling and Layer Migration for VLM Pipeline Inference Acceleration on Edge Devices
4. Cross-Layer Performance Analysis of Single-GPU Large Language Model Inference

3B — Key-Value Stores

10:30–12:10 • LT301

1. ZNStore: Decomposing the B+Tree Write Path for Zoned Namespace SSDs
2. HGB+-Tree: An Hourglass-like B+-tree for Persistent Memory
3. CmptQuiet: Optimizing Tail Latency and Throughput in LSM-Tree Key-Value Stores via Contention-Aware Compaction
4. Breaking the Single-Entry Bottleneck: Dynamic Write Routing for Heterogeneous Key-Value Stores

3C — Hardware Acceleration

10:30–12:10 • LT701

1. Algorithm–Hardware Co-Design of Spiking Transformers for In-Memory Neuromorphic Edge Vision
2. AIE-HE: Accelerating Client-Side Processing of Homomorphic Encryption on AMD AI Engine
3. RayFST: Accelerating Finite-State Transducer Composition with Ray Tracing Cores
4. A Neuromorphic Dual-Stage Swin Transformer for Wetland Classification on Edge Platforms

3D — Graph Learning

10:30–12:10 • LT702

1. ATLAS: Enabling an Asynchronous State-Aware Pipeline for Efficient Training of Temporal Graph Neural Networks
2. FAST: A Holistic Framework for Optimizing Memory-I/O, Computation, and Sampling in Temporal GNN Training
3. PruneInfer: Exact Full-Neighborhood GNN Inference of Large Datasets on a Single GPU via Full Topology Pruning
4. Tempest: A GPU-Accelerated Engine for Streaming Temporal Random Walks

12:10–13:40 Lunch Function Hall (Level 3)

Wednesday, 30 September Afternoon

13:40–14:40 **Keynote 4 — Li-Shiuan Peh**, National University of Singapore
Parallel Computing on Ultra-Low-Power Chips • AUD302

14:40–15:10 Tea Break Function Hall (Level 3)

4A — LLM Inference and Serving III

15:10–16:50 • AUD302

1. NeuroPrefetcher: Storage-Aware Sparse LLM Inference via Delta Prefetching
2. SSQT: A Hardware-Friendly Fusion Compression Framework of Structured Sparsification and Sensitivity-Driven Quantization for Large-Scale Language Models
3. Online Scheduling of Battery-Aware Speculative Decoding for Energy-Efficient Cloud-Edge Collaborative LLM Inference
4. GroupKV: Hierarchical KV Cache Management for Long-Context Diffusion LLM Inference

4B — Storage Systems

15:10–16:50 • LT301

1. PG-Tune: Prediction-Guided Distributed Storage System Tuning without Repeated Online Benchmarking
2. Toward Continuous Service in Enterprise-Level Distributed Block Storage Systems
3. DdIRT: A deterministic data layout for efficient redundancy transitioning in erasure-coded systems
4. FNR: A Probabilistic Data Repair Scheme for Reed-Solomon Coded Storage Amid Network Fluctuations

4C — Sparse Linear Algebra I

15:10–16:50 • LT701

1. Hierarchical Shared Memory-Aware Optimization for TRSM on GPU Platforms
2. DB-SpMSpV: Dual-View Blocked Sparse Matrix-Sparse Vector Multiplication for Dynamic GPU Workloads
3. AFH-SpMM: Auto-Fit Heterogeneous Block Sparse-Dense Matrix Multiplication on Tensor Core GPUs
4. RSH-SpMM: A Row-Structured Hybrid Kernel for Sparse Matrix-Matrix Multiplication on GPUs

4D — Performance Analysis and Modeling I

15:10–16:50 • LT702

1. Revealing NVIDIA Closed-Source Driver Command Streams for CPU–GPU Runtime Behavior Insight
2. MServerSim: A Modular Design based Simulation Technique for Modern AI Servers
3. StreamTrace: Fast Trace Analysis for Large-Scale Parallel Applications on a Single Node
4. Vadar: Runtime Performance Variance Detection and Diagnosis for Parallel Applications

17:15 **Shuttle bus to the banquet**
Departs from NTU@one-north, 11 Slim Barracks Rise, Singapore 138664

18:30–21:00 **Conference Banquet — The Ballroom @ Mount Faber Peak**
109 Mount Faber Road, Singapore 099203

21:00 **Return shuttle bus**
Departs Mount Faber Peak –> Park Avenue Rochester (conference hotel), 31 Rochester Drive, Singapore 138637

Thursday, 1 October Morning

5A — Memory Management

09:00–10:15 • AUD302

1. Governing Global ASID/PCID Management for QoS Isolation in Shared-Kernel Mixed-Priority Systems
2. MiTDM: Eliminating False Conflicts in Scalable MVCC in Disaggregated Memory
3. RDPart: A reuse-based OS-level cache-partitioning policy for fairness optimization in cloud data centers

5B — AI Acceleration

09:00–10:15 • LT301

1. AccESM: Packed-Total Exact Inference for CPU ESM2 Embedding
2. Pagoda: Time and Energy Rooflines for DNN Workloads on Edge Accelerators Across Power Modes
3. MX-KMeans: Accelerating K-Means Clustering through Microscaling Quantization

5C — Serverless Computing

09:00–10:15 • LT701

1. AeroFlow: Accelerating Serverless Workflows through Serialization-Free WebAssembly Execution
2. MOPAR: A Model Partitioning Framework for Deep Learning Inference Services on Serverless Platforms
3. RSC: Enabling High-Performance RDMA for Serverless Services with Adaptive Network Multiplexing

5D — Parallel Data Processing

09:00–10:15 • LT702

1. Chapel for Parallel AND Distributed GPU Computing: A Case Study with Jaccard Similarity
2. Parallel Prefix Filter Search for Weighted Jaccard Similarity
3. SimPattern: A Novel Byte-Pattern Matching based Chunk Similarity Detection for Post-Deduplication

10:15–10:45 Tea Break Function Hall (Level 3)

6A — AI Caching

10:45–12:00 • AUD302

1. PINCH: Predictive Importance-Sampling-Informed Cache for I/O-Bound DNN Training
2. Edge-Cloud Collaborative RAG with Multi-State Semantic Caching
3. MIGServe: Layout-Aware Multi-Instance GPU Management for Efficient LLM Serving

6B — Learning-Based Scheduling

10:45–12:00 • LT301

1. ReLA: Representation Learning and Aggregation for Scalable Job Scheduling with Reinforcement Learning
2. Scalable RL-Based Online Scheduling Framework for Modular Process Systems
3. BP-MDS: Million-Scale Approximate CVRP in Minutes via Parallel Divide-and-Conquer

6C — Network Algorithms

10:45–12:00 • LT701

1. Scalable Exact Path Selection via Structure-Aware Search for Virtual Payment Channels
2. Asynchronous Dispersion with Optimal Time Complexity
3. S2PAR: Scalable Surface Code Ancilla Routing Card for Distributed Fault-Tolerant Quantum Computing

6D — Performance Analysis and Modeling II

10:45–12:00 • LT702

1. Async-Checkpoint: Redesigning Gradient Checkpointing with Asynchronous Execution
2. Offline Signal Reconstruction and Time-dependent Roofline analysis for Always-on Monitoring
3. GCL-Sampler: Discovering Kernel Similarity for Sampled GPU Simulation via Graph Contrastive Learning

12:00–13:00 Lunch Function Hall (Level 3)

Thursday, 1 October Afternoon

7A — Sparse Linear Algebra II

13:00–14:40 • AUD302

1. TFS: Tile-Aware SpMM–GeMM Fusion for Accelerating GNN Inference on Intel AMX
2. CoTC-SpMM: A Cooperative Tensor–CUDA Cores Scheme for Efficient Sparse Matrix Multiplication
3. RODIS: Accelerating Sparse Matrix Multiplication on GPU via Row-Orchestration and Dynamic Instruction Scheduling
4. TileSpMM: A Variable-Size Tiled Algorithm for Sparse Matrix-Matrix Multiplication on Tensor Cores

7B — Data-Center Networks

13:00–14:40 • LT301

1. ActiveRDMA: SmartNIC-offloaded, Multi-Step RDMA operations
2. Coflow Scheduling in Hybrid-Switched Data Center Networks under Not-All-Stop Reconfiguration
3. PipeTree: Decision Tree Partitioning for Efficient Inference on Programmable Switches
4. Bypass-Enabled Topology-Agnostic Multicast Router for Fault-Tolerant Network-on-Chips

7C — Distributed Learning

13:00–14:40 • LT701

1. Low-bit and Sparsified Gradient Communication for Accelerating Distributed Deep Learning with Convergence Guarantees
2. Decentralized Learning with Communication-Efficient Learned Gradient Sketches
3. DiCTR: A Distributed Lightweight Second-Order Optimization Algorithm for Deep Model Training
4. Towards High-Fidelity yet Low-Overhead Gradient Compression for In-Network Aggregation

7D — Parallel Training

13:00–14:40 • LT702

1. OmniPipe: Efficient, Flexible and Scalable Pipeline Parallelism for Large Model Training
2. Accelerating Long-Tail Generation in Synchronous RLHF Training via Adaptive Tensor Parallelism
3. Memory-Aware Joint Optimisation of Partitioning and Scheduling for Pipeline-Parallel Training
4. Chameleon: Adaptive Bidirectional Pipeline Parallelism for Edge Collaborative Training

14:40–15:10 Tea Break Function Hall (Level 3)

8A — GEMM Optimization

15:10–16:50 • AUD302

1. TileGEMM: Boosting the Performance of GEMM on AMX-Powered CPUs by Exploiting Data Reuse
2. BAG: Faster Matrix Multiplication on a Single GPU
3. MAGSAT: Memory-Access-Guided Fusion and Shape-Adaptive Auto-Tuning for Dynamic Tensor Programs
4. RapidGEMM: An Efficient Tile-agnostic GEMM for Mixture-of-Experts Training

8B — Cloud Scheduling

15:10–16:50 • LT301

1. QMScaler: A QoS-Constrained Resource-Efficient Microservice Autoscaling Framework for Edge Environments
2. Microservices Scheduling with ReS Model: Characterization, Regulation and Metrics
3. FaaSHive: Addressing Limited Visibility in Function-as-a-Service via Worker-driven Scheduling

8C — Federated Learning

15:10–16:50 • LT701

1. SCPFL: A Safety Reinforcement Learning Framework for Joint Energy-Fairness Optimization in Long-term Federated Learning
2. Scaling Synthetic-Image Pre-Training for Federated Fine-Tuning of Large Vision Models
3. Tensos: Fast and Accurate Federated GBDT Training via Tentative Feature Shrinking on Stragglers

8D — Scientific Computing II

15:10–16:50 • LT702

1. MillionQAOA: A Dividable QAOA Framework for Solving Million-Vertex-Scale Max-Cut Problems
2. Bit-Efficient Tensor Core Acceleration for High-Order Shallow-Water Dynamical Core
3. A Low-Overhead, Lightweight, Lossless GPU Compression Algorithm for Numerical Simulations

16:50 Conference closes

Conference Venue

NTU@one-north
11 Slim Barracks Rise
Singapore 138664

Nearest MRT

Buona Vista (EW21 / CC22), **Exit D**
7–8 minutes on foot to the venue

CONFERENCE FLOORS

Enter via the **Alumni House entrance on Level 3** and follow the ICPP signs to the registration desk. Executive Centre **Lift Lobby 2** connects Level 3 and Level 7.

Level 3 Keynotes, A and B sessions

AUD302 Auditorium	**Keynotes, A sessions** From registration, follow signs toward the auditorium side of the floor.
LT301 Lecture Theatre	**B sessions** Follow signs toward the Executive Centre / Lift Lobby 2 side.

Level 7 C and D sessions, workshops, tutorials

LT701 **LT702** Lecture Theatres	**C sessions (LT701), D sessions (LT702)** LT701 is the first theatre along the upper corridor after the lift; continue for LT702.
SR703 **SR706** Seminar Rooms	**Workshops and tutorials** SR706 (and SR706A) along the lower central corridor; SR703 and SR704 at the far end of the floor.

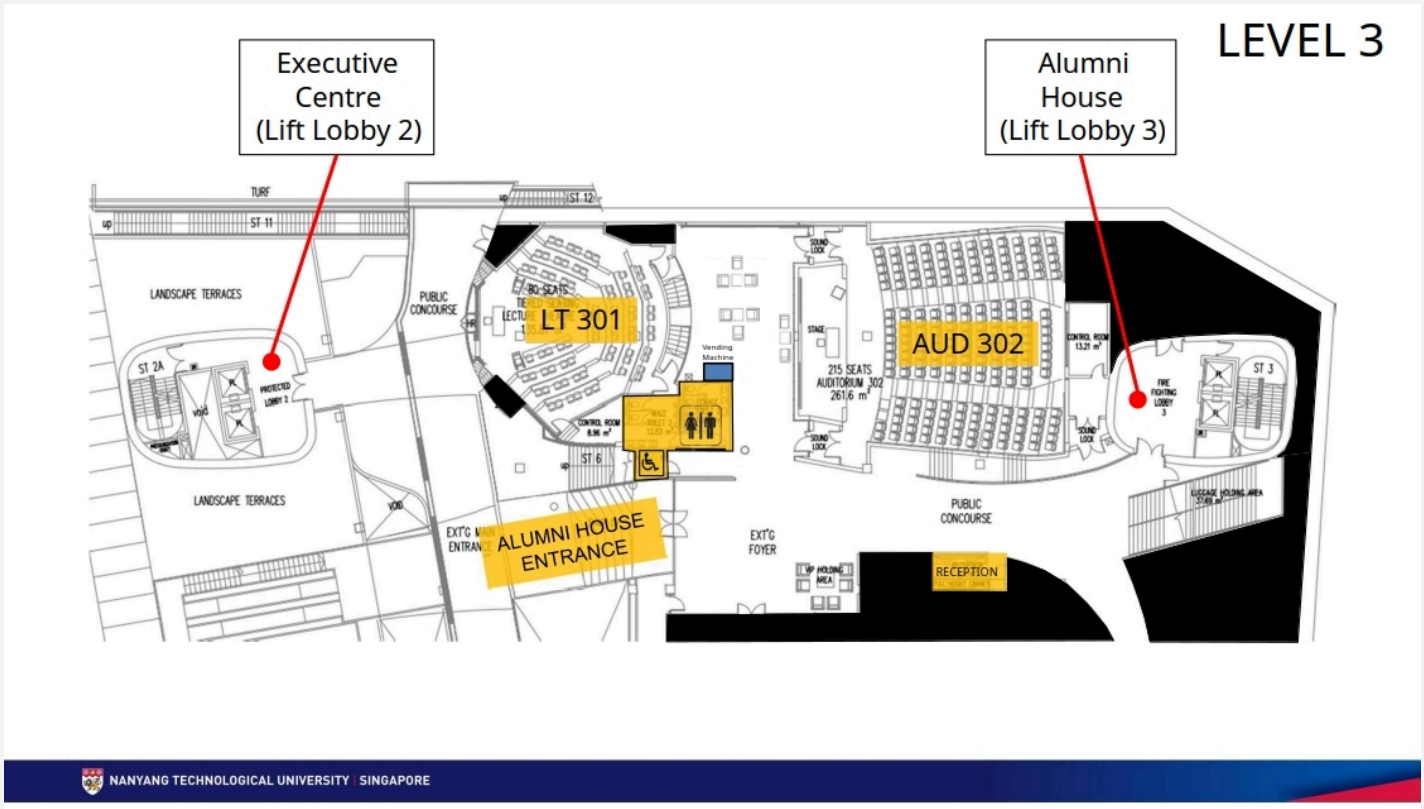
FROM THE CONFERENCE HOTELS

Park Avenue Rochester 31 Rochester Drive, Singapore 138637	8–10 minutes on foot. Walk toward North Buona Vista Road / Buona Vista MRT, cross toward the MOE / one-north Park side, then follow the pedestrian path to Slim Barracks Rise.
Citadines Connect Rochester 1 Rochester Park, Singapore 139212	8–10 minutes on foot. Walk toward Buona Vista MRT / North Buona Vista Road, cross toward one-north Park, and continue along the pedestrian route to Slim Barracks Rise.
Citadines Science Park 7 Science Park Drive, Singapore 118223	Walk 5–10 minutes to Kent Ridge MRT (CC24), take the Circle Line two stops to Buona Vista (CC22), leave by Exit D and walk 7–8 minutes to the venue.

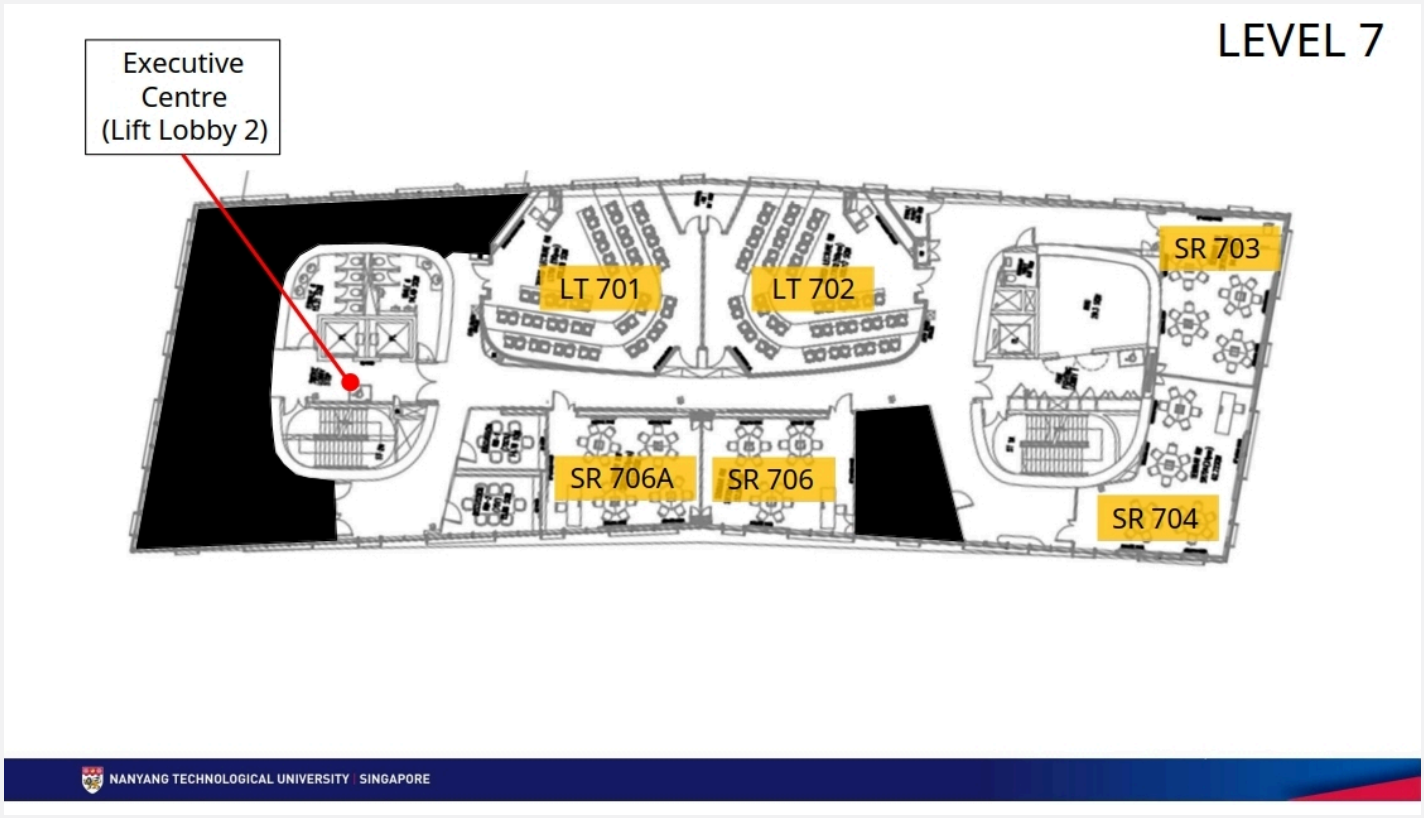
SIGHTSEEING IN SINGAPORE

Scan for our recommended sights, dining spots and getting-around tips.
icpp2026.github.io/social-special-events/#sightseeing

NTU@one-north Floor Plans



Level 3 Registration, AUD302, LT301



Level 7 LT701, LT702, SR703, SR704, SR706, SR706A

CONFERENCE VENUE

NTU@one-north

11 Slim Barracks Rise, Singapore 138664
Buona Vista MRT (EW21 / CC22) under 10 min on foot
Enter via Alumni House, Level 3

Conference floors

Level 3: AUD302, LT301
Level 7: LT701, LT702, SR703, SR704, SR706, SR706A
Executive Centre Lift Lobby 2 connects both floors

REGISTRATION & SOCIAL EVENTS

Registration

Main conference: NTU@one-north, Level 3, outside the Function Hall
Workshops: NTU@one-north, Level 7

Reception

Tue 29 Sep 18:00–20:00
Function Hall, Level 3, NTU@one-north

Banquet

Wed 30 Sep 18:30–21:00
The Ballroom @ Mount Faber Peak, 109 Mount Faber Road

17:15 Shuttle NTU@one-north –> Mount Faber Peak

21:00 Shuttle Mount Faber Peak –> Park Avenue Rochester

ESSENTIALS

Wi-Fi

Two ways to connect. Use **either** one.

Option 1 – Scan the QR code

Only if you have a Singapore-registered mobile number. Follow the instructions on the page to register.

www3.ntu.edu.sg/cits2/wireless/smsguide/smsguide.htm

Option 2 – Use a conference account

For everyone else. Network: **NTUSECURE**. Log in with one of these:

LOGIN ACCESS 1 (AUD302)
ID: **onec023**
PW: **onecuser023**

LOGIN ACCESS 2 (AUD302)
ID: **onec024**
PW: **onecuser024**

PLATINUM SPONSOR

HUAWEI

IN COOPERATION WITH

acm In-Cooperation

sighpc

ORGANIZED BY

IACC
International Association for Computers and Communications

SOCIETY OF GREEN AND SUSTAINABLE COMPUTING SINGAPORE