



UNIVERSITY OF  
MAURITIUS

## Thursday 20th August 2026

16:00 - 16:30 Welcome and Opening

### 16:30 - 18:00 Recorded Research Session 1: LLM Inference & Serving

#### **KVServe: Service-Aware KV Cache Compression for Communication-Efficient Disaggregated LLM Serving**

Zedong Liu, Xinyang Ma, Dejun Luo (University of Chinese Academy of Sciences), Hairui Zhao (Institute of Computing Technology, Chinese Academy of Sciences), Bing Lu (Institute of Computing Technology Chinese Academy of Sciences), Wenjing Huang (University of Chinese Academy of Sciences), Yida Gu, Xingchen Liu, Zheng Wei (Institute of Computing Technology, Chinese Academy of Sciences), Jinyang Liu (University of California, Riverside), Dingwen Tao, Guangming Tan (Institute of Computing Technology, Chinese Academy of Sciences)

[Link to Paper](#)

#### **DualPath: Accelerating Agentic LLM Inference by Harvesting Disaggregated KV-Cache Storage I/O**

Yongtong Wu (Peking University, DeepSeek-AI), Shaoyuan Chen (Tsinghua University, DeepSeek-AI), Rilun Huang (Peking University), Yixuan Tan (DeepSeek-AI), Yinmin Zhong (Peking University, DeepSeek-AI), Mingxing Zhang (Tsinghua University), Xin Jin (Peking University), Panpan Huang (DeepSeek-AI)

[Link to Paper](#)

#### **Efficient Remote KV Cache Reuse with GPU-native Video Codec**

Liang Mi (Nanjing University and Institute for AI Industry Research (AIR), Tsinghua University), Weijun Wang (Institute for AI Industry Research (AIR), Tsinghua University), Jinghan Chen (Nanjing University), Ting Cao (Institute for AI Industry Research (AIR), Tsinghua University), Haipeng Dai (Nanjing University), Yunxin Liu (Institute for AI Industry Research (AIR), Tsinghua University)

[Link to Paper](#)

#### **Connex: Endpoint Mobility Primitives for Dynamic LLM Serving**

Yanying Lin (University of Chinese Academy of Sciences, Penn), Vincent Liu, Tao Luo (University of Pennsylvania), ChengZhong Xu (University of Macau), Kejiang Ye (Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences)

[Link to Paper](#)

#### **TurboBus: Pooling PCIe Bandwidth for LLM Workloads via Scale-Up Fabrics**

Xinyu Yang, Kaiqiang Xu, Kai Chen (Hong Kong University of Science and Technology)

[Link to Paper](#)

## Thursday 20th August 2026

|               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|---------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 18:00-18:15   | Break                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| 18:15 - 19:15 | <b>Recorded Experience Session 1: AI Datacenter Networks &amp; Communication</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|               | <b>Connecting 100K+ GPUs: Building the Communication Stack for Large-Scale LLM Training</b><br>Hongyi Zeng, Min Si, Pavan Balaji, Yongzhou Chen, Ching-Hsiang Chu, Adithya Gangidi, Prashanth Kannan, Bingzhe Liu, Saif Hasan, Dong He, Deep Shah, Ashmitha Jeevaraj Shetty, Gregory R Steinbrecher, Srikanth Sundaresan, Yulun Wang, Yexin Wu, Mingran Yang, Kenny Yu, Minlan Yu (Harvard University), Cen Zhao, Shengbao Zheng, Wesley Bland, Denis Boyda, Suman Gumudavelli, Subodh Iyengar, Daniel Johnson, Cristian Lumezanu, Kai Luo, Rui Miao, Zhe Qu, Venkatraghavan Ramesh, Jingliang Ren, Maxim Samoylov, Jan Seidel, Qiye Tan, Feng Tian, Xinfeng Xie, Jingyi Yang, Yimeng Zhao, Shuqiang Zhang, Tiane "Art" Zhu (Meta)<br><a href="#">Link to Paper</a> |
|               | <b>DistDPU: A Disaggregated DPU Architecture for High-Performance and Cost-Efficient AI Clouds</b><br>Hao Mei (Fudan University), Lizhou Gao (Tencent), Yuanyi Zhu (Fudan University), Liang Wang, Peng Yang, Chao Pei (Tencent), Chuhao Chen, Zijian Li (Fudan University), Jian Zhao, Dongbo Gu, Hongchen Ren, Jiyuan Chen, Junpeng Zhang, Yunpeng Guan, Jianye Yuan, Jian Wang (Tencent), Yibo Huang (University of Michigan), Yang Xu (Fudan University)<br><a href="#">Link to Paper</a>                                                                                                                                                                                                                                                                       |
|               | <b>Pegasus: A Data Center Network for Bare-Metal AI Cloud</b><br>Xianneng Zou (Tsinghua), Yadong Liu (Tencent), Yiran Zhang (Unaffiliated), Jian Zhao, Zhaoxun Zhou (Tencent), Qing Wang (Nanjing University), Xingyi Li, Liang Wang (Tencent), Xingda Wei (Shanghai Jiao Tong University), Xiaojie Huang, Zhaohe Chen, Yinben Xia (Tencent), Yuxi Wang (Unaffiliated), Kai Ren, Lizhou Gao, Jiajun Liang, Jian Wang, Chunxu Zhao, Weifeng Zhang, Lu Shi, Xin Li, Jiewei Yang, Hao Lu, Yunpeng Guan, Dongbo Gu, Chao Pei, Xiang Li, Zekun He (Tencent), Yong Jiang (Tsinghua), Yachen Wang (Tencent)<br><a href="#">Link to Paper</a>                                                                                                                               |
| 19:15 - 20:00 | Dinner and Networking                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |

## Friday 21st August, 2026

|               |                                                                                                                                                                                                                                                                                                                                     |
|---------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 15:00 - 17:00 | <b>Recorded Experience Session 2: Content Delivery &amp; Video Streaming</b>                                                                                                                                                                                                                                                        |
|               | <b>Prefetching for Short Video Streaming: Experiences from a Longitudinal Evolution at Planetary Scale</b><br>Yinjie Zhang, Ying Chen, Yuming Hu, Aoyang Zhang, Pengcheng Chen, Zhixiang Luo, Shaopeng Li, Zhendong Zhong, Haiqing Tao, Lan Xie, Shenglan Huang, Zhiwei Fan, Feng Qian (ByteDance)<br><a href="#">Link to Paper</a> |

Friday 21st August, 2026

**Adaptive Bitrate Live Streaming over HTTP-FLV: A Practical System Perspective**

Le Zhang, Tong Meng (ByteDance), Bingcong Lu (SJTU and ByteDance), Jinghao Yuan, Lu Chen, Huanting Liu, Lei Xiao, Nailiang Wu, Zhou Sha, Changqing Yan, Jianrong Zhang, Jianxin Kuang (ByteDance), Li Song (SJTU)

[Link to Paper](#)

**Horizon: A Hyper-Edge Observability Engine for Live Streaming Networks**

Ziqian Liu, Daqian Ding (The University of Hong Kong), Rui Han (ByteDance), Shixian Guo (ByteDance), Zhendong Xie, Aifang Xu, Changqian Wang, Kefei Liu (ByteDance), Jialin Li (National University of Singapore), Yunming Xiao (The Chinese University of Hong Kong, Shenzhen), Heming Cui, Yiming Qiu (The University of Hong Kong)

[Link to Paper](#)

**CacheFlare: Optimizing Cold Content Performance in CDNs**

Tiansheng Zhang, Yuling Chen, Ahmed Kamal, Liang Zhou, Jianfeng Tang, Huapeng Zhou, Thilan Ganegedara, Yuhao Guo, Sanjay Sane, Ben Vallis (Meta Platform Inc.), Theophilus A. Benson (Carnegie Mellon University), Ying Zhang (Meta Platform Inc.)

[Link to Paper](#)

**Recorded Experience Session 4: Cloud Data Planes & Network Virtualization**

**Rules Offload Engine (ROE): Accelerating Host SDN Policy Evaluation**

Anshuman Verma, Tian Tan, Ahmed Abdelsalam, Milan Dasgupta, Jonathan Hunter, Zach Libby, Narayanan Ravichandran, Harish Srinivasan, Matt Reat, Nadeen Gebara, Vishal Gondaliya (Microsoft), Ezz Hamed (Meta), Rahul Garlapati, Lok Chand Koppaka, Abdullah Mughrabi, Dev Desai, Alexander Malyshev, Shwetha Bhat, Rohan Kandi, Megan Sng (Microsoft), Tushar Garg (Meta), Muluken Hailesellasie, Andrew Putnam, Derek Chiou, Osman Ertugay (Microsoft), Alireza Dabagh (Microsoft Azure), Vivek Bhanu, Daniel Firestone (Microsoft)

[Link to Paper](#)

**Spillway: Orchestrating DPU and Host into a Unified vSwitching Fabric**

Xiaochong Jiang (Zhejiang University), Dian Fan, Yilong Lv, Naixuan Guan (Alibaba Cloud), Qiming Zhao, Sihan Fu (Zhejiang University), Ye Yang, Yu Zhou, Xuyang Ge, Denghui Wu, Yibin Shen, Guochun Hong, Zhipeng Zhang, Yijian Dong, Yiquan Chen, Shaoliang An, Zhixiong Guo, Yixin Xie, Yisong Qiao, Hongwei Ding, Haitao Jiang, Yimin Liu, Tao Shen, Baosheng Wang, Shize Zhang, Bowen Yang, Zikang Chen, Xiaomin Wu, Rong Wen, Yang Song, Jun Liang, Zhigang Zong, Xing Li (Alibaba Cloud), Chengkun Wei (Zhejiang University), Shunmin Zhu (Alibaba Cloud), Wenzhi Chen (Zhejiang University)

[Link to Paper](#)

**From Nimitz to NetPila: The Evolution of Production-Scale Container Network**

Jiaqi Gao, Chao Qin, Sheng Cheng, Jiamin Cao, Guodong Yang, Zhenyu Zhang, Shuhong Zhu, Ennan Zhai, Dennis Cai (Alibaba Group)

[Link to Paper](#)

17:00-17:20

Break

## Friday 21st August, 2026

|               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|---------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 17:20-19:00   | <b>Recorded Research Session 14: Security, Privacy &amp; Networked Systems</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|               | <b>Towards High-Performance Intrusion Detection with Robustness Guarantees on Programmable Switches at ISP Scale</b><br>Han Zhang, Xuefeng Liu, Linqiang Qian (Tsinghua University), Guyue (Grace) Liu (Peking University), Tianyu Zhang, Kaiyang Zhao, Yantu Tong, Zeji Xiao (Tsinghua University), Dongbiao He (Sangfor Technologies Inc), Yahui Li (Tsinghua University), Ke Ruan (China Telecom), Jilong Wang (Tsinghua University), Yongqing Zhu (Institute of Network Technology & China Telecom Corporation Limited Research Institute), Xia Yin (Tsinghua University)<br><a href="#">Link to Paper</a> |
|               | <b>Zero-Knowledge Cloud Analytics</b><br>Zeyang Zhu, Clarence Lam, Alexander Frolov, Ian Miers, Zaoxing Liu (University of Maryland)<br><a href="#">Link to Paper</a>                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|               | <b>DeepSFU: Scalable Deepfake Detection for Video Conferencing</b><br>Tuan Tran, Shirin Ebadi, S. M. H. Hosseini, Woongsub Shin, Evan Ram, Youngwook Son (University of Colorado Boulder), Seyeon Kim (Korea University), Nam Bui (University of Colorado Denver), Kyunghan Lee (Seoul National University), Eric Keller, Sangtae Ha (University of Colorado Boulder)<br><a href="#">Link to Paper</a>                                                                                                                                                                                                         |
|               | <b>Integrating 2PC with Consensus for Fast Replication</b><br>Yan Chen (SJTU & Tsinghua University), Xinyi Yu, Shengyun Liu, Ruofan Xiong, Tianjing Xu (SJTU), Yongwei Wu (Tsinghua University), Yiming Zhang (SJTU)<br><a href="#">Link to Paper</a>                                                                                                                                                                                                                                                                                                                                                          |
|               | <b>Tuning into the Web: A Low-Cost Access Solution for Developing Countries</b><br>Ayush Pandey, Rohail Asim (New York University Abu Dhabi (NYUAD)), Jean Louis K. E. Fendji (University of Ngaoundere), Talal Rahwan (New York University Abu Dhabi), Matteo Varvello (Nokia Bell Labs), Yasir Zaki (New York University Abu Dhabi)<br><a href="#">Link to Paper</a>                                                                                                                                                                                                                                         |
| 19:00 - 20:00 | Dinner and Networking                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |