更新于 2026 年 11 月 23 日

张明喆

蚂蚁集团蚂蚁技术研究院计算系统实验室副主任

mingzhe-zhang@outlook.com | 个人主页 | ORCID

研究方向

研究聚焦计算机体系结构与系统，主要方向包括同态加密加速系统、新型非易失存储器及系统、领域特定加速器架构、片上网络和硬件安全。

任职与科研经历

- 2024.05 至今 | 蚂蚁集团蚂蚁技术研究院计算系统实验室，副主任。
- 2021.04–2024.05 | 中国科学院信息工程研究所，副研究员。
- 2018.07–2021.04 | 中国科学院计算技术研究所，助理研究员。
- 2015.11–2017.02 | 芝加哥大学，访问学者。合作导师：Frederic T. Chong。
- 2010.10–2013.06 | 中国科学院计算技术研究所，科研实习生。
- 2008.07–2009.09 | 中国科学院软件研究所，工程师。

教育经历

- 2013–2018 | 中国科学院计算技术研究所，计算机科学博士。
- 2009–2013 | 内蒙古大学，计算机科学硕士。
- 2004–2008 | 南京邮电大学，计算机科学工学学士。

科研基金

- 国家自然科学基金青年科学基金（62002339）：面向高吞吐应用场景的相变存储器优化技术研究。2021–2023 年 | 24 万元 | 主持。
- 中国科学院战略性先导科技专项（B 类，XDB44000000）：基于存内计算架构的图计算系统软件栈与硬件算子。2020–2024 年 | 468 万元 | 子课题负责人。
- 计算机体系结构国家重点实验室创新项目：基于器件特性的非易失存储器优化。2020–2021 年 | 5 万元 | 主持。
- 中国科学院战略性先导科技专项培育项目（XDPB12）：基于存内计算的图计算加速。2018–2019 年 | 40 万元 | 子课题共同负责人（负责人：Xiaochun Ye）。

荣誉与奖励

- 2023 | ACM SIGOPS ChinaSys 新星奖。
- 2020 | 阿里巴巴–CCF TCIST 优秀论文奖。
- 2019 | CCF 优秀博士学位论文奖提名。

学术服务

- 会议组织: ISCA 2026 宣传与社交媒体联合主席。曾任 ISCA、MICRO (2025), NPC (2020), NAS、NVMSA（2019）分会主席。
- 程序委员会: ISCA（2025–2027）、MICRO（2025–2026）、HPCA（2024、2027）、CASES (2026)、ICCD（2021–2026）、ICS（2024）、ICPADS（2021）、NVMSA（2019–2024）、NPC (2020)、NAS（2019）、ChinaSys（第 20、24 届）。
- 外部评审委员会: ISCA（2022、2024）、HPCA（2023）、MICRO（2022）。
- 期刊审稿: IEEE TC、TVLSI、TCAD、IEEE Micro、ACM TODAES、TECS、ACM Computing Surveys、《计算机科学技术学报》、《计算机研究与发展》及 JCN。

论文成果

论文索引 | Google Scholar

会议论文

C58 Lei Chen, Yiming Sun, Erci Xu, Yi Chen, Shengyu Fan, Xianglong Deng, Guang Fan, Liang Kong, Yilan Zhu, Geng Yang, Fu Xiao, Shoumeng Yan, Yisong Chang, **Mingzhe Zhang**. HELM: A Scalable Multi-GPU System for Homomorphic LLM Inference. *HPCA 2027*（已录用）.

C57 Tiantian Lin, Tengjun Long, **Mingzhe Zhang**, Yingtao Jiang, Xiaohang Wang, Chao Zhang, Amit Kumar Singh, Jieming Yin, Kui Ren. FCS: Frontier-based Commit-time Synchronization Method for Modular Simulators. *HPCA 2027*（已录用）.

C56 Yijing Ning, Jiankuo Dong, Tian Zhou, Fangyu Zheng, Guang Fan, **Mingzhe Zhang**, Jingqiang Lin. FORGE: FAFFT-Optimized Ring Arithmetic for GPU Execution of HQC and BIKE. *HPCA 2027*（已录用）.

C55 Jianyu Zhao, Zhiyuan Zhang, Guang Fan, **Mingzhe Zhang**, Shoumeng Yan, Lei Ju, Zhuoran Ji. Accelerating Fully Homomorphic Encryption for Large Language Model Inference on Multi-GPU Systems. *PPoPP 2027*（已录用）.

C54 Bo Zhang, Xiaoyi Ye, Mingyu Gao, **Mingzhe Zhang**, Shoumeng Yan. Investigation of Karatsuba Algorithm for Large-Bitwidth Modular Multiplication. *ASP-DAC 2027*（已录用）.

C53 Tianxiang Sui, Qing Zhu, JianXin Lai, **Mingzhe Zhang**, Yan Liu, Long Li, Peng Yuan, Linjie Xiao, Xiaojing Zhang, Shoumeng Yan, Jingling Xue. HPAO: Polynomial-Level Optimization for RNS–CKKS Programs. *CGO 2027*（已录用）.

C52 Jiangrui Yu, Baosheng Zhang, Liang Kong, Lin Ding, Yi Chen, Ye Yu, **Mingzhe Zhang**, Meng Li. ROSETTA: Efficient and Accurate Privacy-Preserving LLM Decoding via Hybrid CKKS/TFHE Evaluation. *CCS 2026*.

C51 Yi Chen, Ziyu Tang, Chao Yang, Guang Fan, **Mingzhe Zhang**, Meng Li. Helios: Melting Kernel Boundaries for GPU-Accelerated HE via Graph Rewriting and Microarchitecture-Aware Mapping. *MICRO 2026*.

C50 Xiurui Pan, Guoci Chen, Guangyu Sun, **Mingzhe Zhang**, Jie Zhang. LUTHE: Breaking the Memory Wall of High-Precision Encrypted LLM Nonlinearities via GPU–CSD Co-Design. *MICRO 2026*.

C49 Tenghui Hua, Huawei Li, Jianan Mu, Jing Ye, Liang Kong, **Mingzhe Zhang**. Chimera: A Unified FHE Accelerator with Enhanced Compatibility for TFHE. *DAC 2026*.

C48 Junyi Zhang, Xianglong Deng, Yi Chen, Guang Fan, Lei Chen, Dian Jiao, Shengyu Fan, Zhiwei Wang, **Mingzhe Zhang**. MNEMOS: A GPU-based TFHE Acceleration Framework with Memory Access Optimization. *ISCA 2026*.

C47 Guang Fan, Yi Chen, Lei Chen, Liang Kong, Chao Niu, Dian Jiao, Yilan Zhu, Geng Yang, Shengyu Fan, Xianglong Deng, Fangyu Zheng, Jian Weng, Yisong Chang, Meng Li, Shoumeng Yan, **Mingzhe Zhang**. Hierarchical Exploitation of Memory Efficiency for GPU-Based FHE Acceleration. *ISCA 2026*.

C46 Zihao Yang, Dian Jiao, Shengyu Fan, Xianglong Deng, Rui Hou, **Mingzhe Zhang**. A Bitwidth-Flexible Modular Multiplier with Shift-Free Accumulation for Efficient NTT Acceleration in FHE. *ISCAS 2026*.

C45 Zhuoran Ji, Jianyu Zhao, Guang Fan, **Mingzhe Zhang**, Shoumeng Yan, Xueyu Wu, Lei Ju. A Framework for Developing and Optimizing Fully Homomorphic Encryption Programs on GPUs. *ASPLOS 2026*.

C44 Song Bian, Yintai Sun, Zian Zhao, Haowen Pan, **Mingzhe Zhang**, Zhenyu Guan. Libra: Pattern-Scheduling Co-Optimization for Cross-Scheme FHE Code Generation over GPGPU. *Security 2026*.

C43 Yuda An, Shushu Yi, Bo Mao, Qiao Li, **Mingzhe Zhang**, Diyu Zhou, Ke Zhou, Nong Xiao, Guangyu Sun, Yingwei Luo, Jie Zhang. Xerxes: Extensive Exploration of Scalable Hardware Systems with CXL-Based Simulation Framework. *FAST 2026*.

C42 Liang Kong, Shengyu Fan, Xianglong Deng, Lei Chen, Guang Fan, Yilan Zhu, Geng Yang, Yisong Chang, Shoumeng Yan, **Mingzhe Zhang**. Falcon: Algorithm-Hardware Co-Design for Efficient Fully Homomorphic Encryption Accelerator. *ASPLOS 2026*.

C41 Yilan Zhu, Geng Yang, Xingyu Tian, Dilshan Kumarathunga, Liang Kong, Xianglong Deng, Shengyu Fan, Guang Fan, Guiming Shi, Lei Chen, Bo Zhang, Yisong Chang, Shoumeng Yan, Zhenman Fang, **Mingzhe Zhang**. An Efficient and Scalable Hardware Architecture for Number Theoretic Transform on FPGA with Design Automation. *HPCA 2026*.

C40 Tianxiang Sui, Jianxin Lai, Long Li, Peng Yuan, Yan Liu, Qing Zhu, Xiaojing Zhang, Linjie Xiao, **Mingzhe Zhang**, Jingling Xue. FHEFusion: Enabling Operator Fusion in FHE Compilers for Depth-Efficient DNN Inference. *CGO 2026*.

C39 Zihao Yang, Yilan Zhu, Rui Hou, Dan Meng, Shengyu Fan, **Mingzhe Zhang**. Analysis of Bit-Flip Attacks on Encrypted Neural Networks. *ICPADS 2025*.

C38 Guiming Shi, Yuchen Wei, Shengyu Fan, Xianglong Deng, Liang Kong, Xianbin LI, Jingwei Cai, Shuwen Deng, **Mingzhe Zhang**, Kaisheng Ma. WPC: Weight Plaintext Compression for CNN Inference based on RNS-CKKS. *CCS 2025*.

C37 Zewei Lai, Jinhui Ye, Xiaohang Wang, Zheang Fu, Amit Kumar Singh, Yingtao Jiang, Kui Ren, Mei Yang, Sihai Qiu, Xiaodong Li, Xin Tang, Jie Song, **Mingzhe Zhang**. On Optimizing Intra- and Inter-chiplet Interconnection Networks in Multi-chiplet Systems for Accelerating FHE Encrypted Neural Network Applications. *CASES 2025*.

C36 Tiantian Lin, Cheng Qiu, Xiaohang Wang, Ling Wang, Zhulin Zheng, Yingtao Jiang, Amit Kumar Singh, Jieming Yin, Sihai Qiu, Xiaodong Li, Xin Tang, Jie Song, **Mingzhe Zhang**, Kui Ren. LEGOSim: A Unified Parallel Simulation Framework for Multi-chiplet Heterogeneous Integration. *MICRO 2025*.

C35 Liang Kong, Shengyu Fan, Xianglong Deng, Guang Fan, Lei Chen, Guiming Shi, Yilan Zhu, Geng Yang, Shoumeng Yan, **Mingzhe Zhang**. HAWK: Fully Homomorphic Encryption Accelerator with Fixed-Word Key Decomposition Switching. *MICRO 2025*.

C34 Lei Chen, Erci Xu, Yiming Sun, Shengyu Fan, Xianglong Deng, Guiming Shi, Guang Fan, Liang Kong, Yilan Zhu, Shoumeng Yan, **Mingzhe Zhang**. The Bleak Future of Fully Homomorphic Encryption: from a Storage I/O Perspective. *APPT 2025*.

C33 Li Peng, Wenbo Wu, Shushu Yi, Xianzhang Chen, Chenxi Wang, Shengwen Liang, Zhe Wang, Nong Xiao, Qiao Li, **Mingzhe Zhang**, Jie Zhang. XHarvest: Rethinking High-Performance and Cost-Efficient SSD Architecture with CXL-Driven Harvesting. *ISCA 2025*.

C32 Dian Jiao, Xianglong Deng, Zhiwei Wang, Shengyu Fan, Yi Chen, Rui Hou, **Mingzhe Zhang**. Neo: Towards Efficient Fully Homomorphic Encryption Acceleration using Tensor Core. *ISCA 2025*.

C31 Shengyu Fan, Xianglong Deng, Liang Kong, Guiming Shi, Guang Fan, Rui Hou, Dan Meng, **Mingzhe Zhang**. FAST:An FHE Accelerator for Scalable-parallelism with Tunable-bit. *ISCA 2025*.

C30 Fangxin Liu, Haomin Li, Zongwu Wang, Bo Zhang, **Mingzhe Zhang**, Shoumeng Yan, Li Jiang. ALLMod: Exploring Area-Efficiency of LUT-based Large Number Modular Reduction via Hybrid Workloads. *DAC 2025*.

C29 Zihao Yang, Mengxin Zheng, Shengyu Fan, Qian Lou, Rui Hou, Dan Meng, **Mingzhe Zhang**. Corrosion Hammer: A Self-Activated Bit-Flip Attack to the Processing-In-Memory Accelerator. *CF 2025*.

C28 Yancheng Zhang, Jiaqi Xue, Mengxin Zheng, Mimi Xie, **Mingzhe Zhang**, Lei Jiang, Qian Lou. CipherPrune: Efficient and Scalable Private Transformer Inference. *ICLR 2025*.

C27 Guang Fan, **Mingzhe Zhang**, Fangyu Zheng, Shengyu Fan, Tian Zhou, Xianglong Deng, Wenxu Tang, Liang Kong, Yixuan Song, Shoumeng Yan. WarpDrive: GPU-Based Fully Homomorphic Encryption Acceleration Leveraging Tensor and CUDA Cores. *HPCA 2025*.

C26 Xianglong Deng, Shengyu Fan, Zhicheng Hu, Zhuoyu Tian, Zihao Yang, Jiangrui Yu, Dingyuan Cao, Dan Meng, Rui Hou, Meng Li, Qian Lou, **Mingzhe Zhang**. Trinity: A General Purpose FHE Accelerator. *MICRO 2024*.

C25 Xiurui Pan, Yuda An, Shengwen Liang, Bo Mao, **Mingzhe Zhang**, Qiao Li, Myoungsoo Jung, Jie Zhang. Flagger: Cooperative Acceleration for Large-Scale Cross-Silo Federated Learning Aggregation. *ISCA 2024*.

C24 Jianan Mu, Husheng Han, Shangyi Shi, Jing Ye, Zizhen Liu, Shengwen Liang, Meng Li, **Mingzhe Zhang**, Song Bian, Xing Hu, Huawei Li, Xiaowei Li. Alchemist: A Unified Accelerator Architecture for Cross-Scheme Fully Homomorphic Encryption. *DAC 2024*.

C23 Yinghao Yang, Huaizhi Zhang, Shengyu Fan, Hang Lu, **Mingzhe Zhang**, Xiaowei Li. Poseidon: Practical Homomorphic Encryption Accelerator. *HPCA 2023*.

C22 Shengyu Fan, Zhiwei Wang, Weizhi Xu, Rui Hou, Dan Meng, **Mingzhe Zhang**. TensorFHE: Achieving Practical Computation on Encrypted Data Using GPGPU. *HPCA 2023*.

C21 Yajuan Du, Mingyang Liu, Yuqi Yang, **Mingzhe Zhang** and Xulong Tang. Enhancing GPU Performance via Neighboring Directory Table Based Inter-TLB Sharing. *ICCD 2022*.

C20 Hang Lu, Liang Chang, Chenglong Li, Zixuan Zhu, Shengjian Lu, Yanhuan Liu, **Mingzhe Zhang**. Distilling Bit-level Sparsity Parallelism for General Purpose Deep Learning Acceleration. *MICRO 2021*.

C19 Hongyan Li, Hang Lu, Jiawen Huang, Wenxu Wang, **Mingzhe Zhang**, Wei Chen, Liang Chang and Xiaowei Li. BitX: Empower Versatile Inference with Hardware Runtime Pruning. *ICPP 2021*.

C18 Liang Yan, **Mingzhe Zhang**, Rujia Wang, Xiaoming Chen, Xingqi Zou, Xiaoyang Lu, Yinhe Han, Xian-He Sun. CoPIM: A Concurrency-aware PIM Workload Offloading Architecture for Graph Applications. *ISLPED 2021*.

C17 Dingyuan Cao, **Mingzhe Zhang**, Hang Lu, Xiaochun Ye, Dongrui Fan, Yuezhi Che, Rujia Wang. Streamline Ring ORAM Accesses through Spatial and Temporal Optimization. *HPCA 2021*.

C16 Chundian Li, **Mingzhe Zhang**, Zhiwei Xu, Xianhe Sun. Self-adaptive Address Mapping Mechanism for Access Pattern Awareness on DRAM. *ISPA 2019*.

C15 Ning Lin, Hang Lu, Jingliang Gao, **Mingzhe Zhang**, Xiaowei Li. When Deep Learning Meets the Edge: Auto-Masking Deep Neural Networks for Efficient Machine Learning on Edge Devices. *ICCD 2019*.

C14 **Mingzhe Zhang**, Lunkai Zhang, Frederic T. Chong, Zhiyong Liu. Balancing Performance and Energy Efficiency of ONoC by Using Adaptive Bandwidth. *ICCD 2019*.

C13 Farzaneh Zokaee, **Mingzhe Zhang**, Lei Jiang. FindeR: Accelerating FM-Index-based Exact Pattern Matching in Genomic Sequences through ReRAM technology. *PACT 2019*.

C12 Shuqian An, **Mingzhe Zhang**, Xiaochun Ye, Da Wang, Hao Zhang, Dongrui Fan, Zhimin Tang. C-MAP: Improving the Effectiveness of Mapping Method for CGRA by Reducing NoC Congestion. *HPCC 2019*.

C11 Farzaneh Zokaee, **Mingzhe Zhang**, Xiaochun Ye, Dongrui Fan, Lei Jiang. Magma: A Monolithic 3D Vertical Heterogeneous ReRAM-based Main Memory Architecture. *DAC 2019*.

C10 Zihao Wang, **Mingzhe Zhang**, Jingrong Zhang, Rui Yan, Xiaohua Wan, Zhiyong Liu, Fa Zhang, Xuefeng Cui. Mmalloc: A Dynamic Memory Management on Many-core Coprocessor for the Acceleration of Storage-intensive Bioinformatics Application. *BIBM 2018*.

C09 **Mingzhe Zhang**, Lunkai Zhang, Lei Jiang, Frederic T Chong, Zhiyong Liu. Quick-and-Dirty: Improving Performance of MLC PCM by Using Temporary Short Writes. *ICCD 2017*.

C08 **Mingzhe Zhang**, Lunkai Zhang, Lei Jiang, Zhiyong Liu, Frederic T Chong. Balancing performance and lifetime of MLC PCM by using a region retention monitor. *HPCA 2017*.

C07 **Mingzhe Zhang**, Yangguang Shi, Fa Zhang, Zhiyong Liu. COMRANCE: A rapid method for Network-on-Chip design space exploration. *IGSC 2016*.

C06 Lunkai Zhang, Dmitri Strukov, Hebatallah Saadeldeen, Dongrui Fan, **Mingzhe Zhang**, Diana Franklin. SpongeDirectory: Flexible sparse directories utilizing multi-level memristors. *PACT 2014*.

C05 Xiaochun Ye, Dongrui Fan, Ninghui Sun, Shibin Tang, **Mingzhe Zhang**, Hao Zhang. SimICT: A fast and flexible framework for performance and power evaluation of large-scale architecture. *ISLPED 2013*.

C04 Lunkai Zhang, **Mingzhe Zhang**, Lingjun Fan, Da Wang, Paolo Ienne. Spontaneous reload cache: Mimicking a larger cache with minimal hardware requirement. *NAS 2013*.

C03 Shuai Zhang, Zhiyong Liu, Dongrui Fan, Fonglong Song, **Mingzhe Zhang**. Energy-Performance Modeling and Optimization of Parallel Computing in On-Chip Networks. *ISPA 2013*.

C02 **Mingzhe Zhang**, Da Wang, Xiaochun Ye, Liqiang He, Dongrui Fan, Zhiyong Liu. A Path-Adaptive Opto-electronic Hybrid NoC for Chip Multi-processor. *ISPA 2013*.

C01 **Mingzhe Zhang**, Liqiang He, Dongrui Fan. Self-Correction Trace Model: A Full-System Simulator for Optical Network-on-Chip. *IPDPSW 2012*.

期刊论文

J22 Yuda An, Shushu Yi, Bo Mao, Qiao Li, **Mingzhe Zhang**, Diyu Zhou, Ke Zhou, Nong Xiao, Guangyu Sun, Yingwei Luo, Jie Zhang. A Cross-Layer Simulation Framework for Protocol Efficiency and Architecture Exploration in Scalable CXL Fabrics. *ACM Transactions on Computer Systems (TOCS)*, 2026.

J21 Shengyu Fan, Xianglong Deng, Liang Kong, Guang Fan, Zhicheng Hu, Liang Chang, Dan Meng, Rui Hou, **Mingzhe Zhang**. Taiyi: An Accelerator Architecture Optimized for Advanced KeySwitch in FHE. *IEEE Transactions on Computers*, 2026.

J20 Xianglong Deng, Shengyu Fan, Zhicheng Hu, Zhuoyu Tian, Aisen Xie, Liang Chang, Meng Li, Rui Hou, Dan Meng, **Mingzhe Zhang**. ArborDorp: An Area Efficient CKKS Accelerator. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2026.

J19 Xianglong Deng, Shengyu Fan, Liang Kong, Yilan Zhu, Geng Yang, Lei Chen, Guang Fan, Yi Chen, Yisong Chang, Jian Weng, Shoumeng Yan, **Mingzhe Zhang**. CACO: Chiplet-based FHE Acceleration with Inter-Chiplet Communication Optimizations. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2026.

J18 Shengyu Fan, Xianglong Deng, Liang Kong, Guang Fan, Lei Chen, Yilan Zhu, Geng Yang, Yisong Chang, Shoumeng Yan, **Mingzhe Zhang**. OmniCifro: A Heterogeneous Acceleration System with PIM for Encrypted Computation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2026.

J17 Bo Zhang, **Mingzhe Zhang**, Shoumeng Yan. Binary Subscript Representation and Efficient Pipelined Number Theoretic Transform. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2026.

J16 Yintai Sun, Shengyu Fan, Zhenhua Yin, Xinkai Song, Xing Hu, Zidong Du, Qi Guo, Weizhi Xu, Rui Hou, Dan Meng, Song Bian, **Mingzhe Zhang**. TensorFHE+: Fully Homomorphic Encryption Acceleration based on Linear Algebra. *IEEE Transactions on Computers, early access*, 2025.

J15 ZIHAO YANG; Mengxin Zheng; Shengyu Fan; Qian Lou; Rui Hou; Dan Meng; **Mingzhe Zhang**. Corrosion Hammer: A Self-Activated Bit-Flip Attack to the Processing-In-Memory Accelerator. *CyberSecurity*, 2025.

J14 Shengyu fan; Xianglong Deng; Xulong Tang; Weizhi Xu; **Mingzhe Zhang**. Understanding and Boosting Fully Homomorphic Encryption Applications on GPU. *CyberSecurity*, 2025.

J13 Zhuoyu Tian; Shengyu Fan; Xianglong Deng; Rui Hou; Dan Meng; **Mingzhe Zhang**. A Survey on Fully Homomorphic Encryption Compilers. *CyberSecurity*, 2025.

J12 Bo Zhang, **Mingzhe Zhang**, Shoumeng Yan. Exploration of Karatsuba Algorithm for Efficient Barrett Modular Multiplication. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2025.

J11 Chenjia Xie, Zhuang Shao, **Mingzhe Zhang**, Yuan Du, Li Du. RAC-NAF: A Reconfigurable Analog Circuitry for Nonlinear Activation Function Computation in Computing-in-Memory. *IEEE Journal of Solid-State Circuits (JSSC), Vol 60, Issue 10, pp. 3738 - 3748 (2025).*

J10 Yucheng Zhang, Wenbin Zeng, Hong Jiang, Dan Feng, Zichen Xu, Shuibing He, **Mingzhe Zhang**, Dan Wu. An Efficient Delta Compression Framework Seamlessly Integrated into Inline Deduplication. *ACM Transactions on Storage (TOS), Vol 21, Issue 4, pp. 1-30 (2025).*

J09 Zhuoyu Tian; Lei Chen; Shengyu Fan; Xianglong Deng; Rui Hou; Dan Meng; **Mingzhe Zhang**. LP-HENN: Fully Homomorphic Encryption Accelerator with High Energy Efficiency. *CyberSecurity, Vol 8, Artical No 98 (2025).*

J08 Mo Zou, **Mingzhe Zhang**, Rujia Wang, Xian-He Sun, Xiaochun Ye, Dong-Rui Fan, Zhimin Tang. Skyway: Accelerate Graph Applications with a Dual-Path Architecture and Fine-Grained Data Management. *Journal of Computer Science and Technology, 39(4): 871-894 (2024).*

J07 Wenxing Li, Ning Lin, **Mingzhe Zhang**, Hang Lu, Xiaoming Chen, Xiaowei Li. VNet: a versatile network to train real-time semantic segmentation models on a single GPU. *Science China Information Sciences, Vol. 64, Issue 3, pp. 1-2*, 2022.

J06 Mo Zou, **Mingzhe Zhang**, Rujia Wang, Xian-He Sun, Xiaochun Ye, Dongrui Fan, Zhimin Tang. Accelerating Graph Processing with Lightweight Learning-Based Data Reordering. *IEEE Computer Architecture Letters, Vol. 21, Issue 1, pp. 5-8*, 2022.

J05 Wenze Zhao, Yajuan Du, **Mingzhe Zhang**, Mingyang Liu, Kailun Jin, Rachata Ausavarungnirun. Application-Oriented Data Migration to Accelerate In-Memory Database on Hybrid Memory. *Micromachines, Vol. 13, Issue 1, pp. 52-60*, 2022.

J04 Hang Lu, **Mingzhe Zhang**, Yinhe Han, Huawei Li, Li Xiaowei. Architecting Effectual Computation for Machine Learning Accelerators. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), Vol. 39, Issue 10, pp. 2654-2667*, 2020.

J03 **Mingzhe Zhang**, Fa Zhang, Zhiyong Liu. A Survey on Architecture Research of Non-Volatile Memory based on Dynamical Trade-off (in chinese). *Journal of Computer Research and Development, Vol. 56, Issue 4, pp. 677-691*, 2019.

J02 **Mingzhe Zhang**, Lunkai Zhang, Lei Jiang, Frederic T Chong, Zhiyong Liu. Quick-and-Dirty: An Architecture for High-Performance Temporary Short Writes in MLC PCM. *IEEE Transactions on Computers (TC), Vol. 68, Issue 9, pp. 1365-1375*, 2019.

J01 Shaoli Liu, Tianshi Chen, Ling Li, Xi Li, **Mingzhe Zhang**, Chao Wang, Haibo Meng, Xuehai Zhou, Yunji Chen. FreeRider: Non-local adaptive network-on-chip routing with packet-carried propagation of congestion information. *IEEE Transactions on Parallel and Distributed Systems (TPDS), Vol. 26, Issue 8, pp. 2272-2285*, 2015.

预印本

i3 Zeng, Wenxuan, Xu, Tianshi, Chen, Yi, Zhou, Yifan, **Zhang, Mingzhe**, Tan, Jin, Hong, Cheng, Li, Meng. Towards Efficient Privacy-Preserving Machine Learning: A Systematic Review from Protocol, Model, and System Perspectives. *arXiv:2507.14519, 2025.*

i2 An, Yuda, Yi, Shushu, Mao, Bo, Li, Qiao, **Zhang, Mingzhe**, Zhou, Ke, Xiao, Nong, Sun, Guangyu, Wang, Xiaolin, Luo, Yingwei, Zhang, Jie. A Novel Extensible Simulation Framework for CXL-Enabled Systems. *arXiv:2411.08312, 2024.*

i1 Xiang, Taoran, Zhang, Lunkai, An, Shuqian, Ye, Xiaochun, **Zhang, Mingzhe**, Liu, Yanhuan, Yan, Mingyu, Wang, Da, Zhang, Hao, Li, Wenming, Sun, Ninghui, Fan, Dongrui. RISC-NN: Use RISC, NOT CISC as Neural Network Hardware Infrastructure. *arXiv:2103.12393, 2021.*