GCCache: A Non-blocking and Low-latency General-purpose Context Cache Design for RDMA NICs

Zhixiang Zhao* **, Zhichuan Guo* **, Mangu Song*, Zhiyuan Ling*

*National Network New Media Engineering Research Center, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

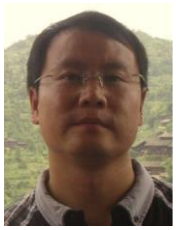
**School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

zhaozx@dsp.ac.cn, guozc@dsp.ac.cn, songmg@dsp.ac.cn, lingzy@dsp.ac.cn

***Abstract*—Remote Direct Memory Access (RDMA) network interface cards (RNICs) are widely used in AI large-scale model training and intelligent computing centers. However, current RNIC caches face several challenges, such as head-of-line (HOL) blocking, repeated processing of misses to the same memory address, high latency, and poor compatibility. To address these challenges, for the first time we propose GCCache: a non-blocking, low-latency, high-performance general-purpose cache architecture for RNICs. GCCache decouples hit processing and miss processing to enable non-blocking cache access and resolve HOL blocking. In addition, an outstanding request management mechanism ensures that for multiple misses to the same memory address, all corresponding requests are served immediately once the context is fetched, thereby saving PCIe bandwidth. Moreover, multi-request parallel processing and pipeline design reduce access latency and support context prefetching. GCCache supports all types of RDMA contexts (QPC, MPT, MTT, etc.) and multi-level page table queries. To the best of our knowledge, GCCache outperforms all currently published RDMA cache architectures in terms of latency and on-chip storage resources, with a cache processing latency of only 5 clock cycles and a throughput of 100 Gbps.**

***Keyword*—RDMA, cache architecture, FPGA.**

Zhixiang Zhao received the B.S. degree in communication engineering from Northeastern University, Qinhuangdao, China, in 2021, and he is currently pursuing the Ph.D. degree at the school of electronic, electrical and communication engineering of the University of Chinese Academy of Sciences. His current research interests include RDMA-based SmartNIC design and FPGA-based network function offloading.



Zhichuan Guo received the B.S. degree from Wuhan University in 1996, and the Ph.D. degree from the University of Science and Technology of China in 2006. From 1996 to 2003, he served as an Electronics Engineer with the 13th Research Institute of China Electronics Technology Group Corporation and a SDH hardware R&D system engineer of optical networks at Huawei. In 2006, he joined with the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China. Now he is a Professor of CAS engaging in field programmable gate array (FPGA)-based code acceleration, VLSI, RDMA and security.

Mangu Song received the M.Sc. degree in Electronics and Communication Engineering from the School of Microelectronics, Chinese Academy of Science, Beijing, China. Currently she is an assistant research fellow in National Network New Media Engineering Research Center, the Institute of Acoustics of Chinese Academy of Sciences. Her research interests include field programmable gate array (FPGA)-based code acceleration and SmartNIC.

Zhiyuan Ling received his B.S. degree in electronic and information engineering from North China Electric Power University in 2018, and his Ph.D. degree in signal and information processing from Institute of Acoustics, Chinese Academy of Sciences (IACAS) in 2023. Currently he is an assistant research fellow in National Network New Media Engineering Research Center, IACAS. His research interests include RDMA, advanced network and SDN technology.