

Bojie Li

Last update on March 10, 2026

bojieli@gmail.com • +86.15011272877 • Haidian, Beijing, China • <https://o1.me>

Work Experience

- Chief Scientist, Pine AI BEIJING, CHINA & SAN JOSE, US
Series A startup (\$25M raised, led by Fortwest) Jan. '25 – present
Leads research on Pine AI's core technology powering autonomous, real-time voice systems that make phone calls and operate computers on behalf of users. Achieved 93% success rate in negotiations with businesses, and saved consumers over \$37M in total.
- Co-Founder & CTO, Stealth Web3 + AI Startup BEIJING, CHINA
Low-Latency and Cost-Effective AI Systems Serving Millions of Web3 Users Aug. '23 – Oct. '24
 - Designed a **NFT generation pipeline** creating \$30M+ value digital assets with generative AI.
 - Designed an **ultra-low latency voice engine** (500ms end-to-end) running on consumer GPUs (400x cheaper than OpenAI) that powers leading virtual character apps like Candy.ai (100K+ paid subscribers).
- Huawei TopMinds Talent Program BEIJING, CHINA
Technical Expert, Computer Networking and Protocol Lab, Beijing HW Digital Technologies June '19 – July '23
 - One of 4 **top-tier Huawei talents** selected from ten thousands of new campus hires in 2019.
 - Key architect of Huawei's **Unified Bus (UB)**, leading a global **team of 25** to build the next-gen datacenter interconnect and **uRPC** unified RPC. Designed an architecture to solve RDMA scalability ($O(n^2) \rightarrow O(n)$) by decoupling transaction/transport layers. uRPC achieved 5–10x performance and 10–100x lower latency compared to gRPC.
 - Led a **team of 15** on **AKG (PLDI '21)**, a tensor compiler for Ascend NPU using polyhedral transformations.
-

Education

- University of Science and Technology of China HEFEI, ANHUI, CHINA
Ph.D. in Computer Science and Technology Sept. '14 – June '19
Joint Ph.D. program with Microsoft Research Asia. Area: Datacenter Systems. Advisors: Prof. Enhong Chen and Dr. Lintao Zhang
- University of Science and Technology of China HEFEI, ANHUI, CHINA
B.S. in Computer Science (School of Gifted Young) Sept. '10 – July '14
-

Internship Experience

- Microsoft Research Asia BEIJING, CHINA
Systems Research Group Oct. '16 – May '19
Advisor: Dr. Lintao Zhang.
 - KV-Direct (SOSP'17)**: Programmable NIC-based key-value store achieving 1.2 billion operations per second (10x faster than CPU).
 - SocksDirect (SIGCOMM'19)**: High-performance user-space socket system compatible with existing apps, improving Nginx performance by 5.5x.
- Microsoft Research Asia BEIJING, CHINA
Wireless and Networking Research Group July '13 – Sept. '14 and July '15 – Oct. '16
Advisor: Dr. Kun Tan.
 - ClickNP (SIGCOMM'16)**: FPGA-based network processing platform achieving 40 Gbps line rate with high-level language.
-

Selected Projects

- Self-Evolving Real-Time Voice Agents PINE AI
Series A startup (\$25M raised, led by Fortwest) Jan. '25 – present
 - Pioneered an agent framework that enables agents to **think while listening** (parallel inference during speech) and **speak while thinking** (streaming response generation) with sub-second latency, significantly outperforming traditional VAD+ASR+LLM+TTS pipelines (2–10s latency).
 - Built self-evolving agent learning framework combining post-training, in-context learning, and externalized learning.

- Built systems-level infrastructure for high performance and cost efficiency, including a **message gateway** orchestrating thousands of concurrent connections across Telegram and WhatsApp.
- Performed holistic optimization across orchestration, voice processing, and computer-use agent components for efficient inter-component communication.

Unified Bus: Next Generation Data Center Interconnect

HUAWEI

Published in SIGCOMM '21 and APNet '23 (first author); with Hunan University: published in ICNP '21, ToN '25, and INFOCOM '25; advised by Dr. Kun Tan

Apr. '20 – July '23

- Key architect of Unified Bus software, leading a team of 25 developers.
- Designed **uRPC**, a high-performance unified RPC for datacenter and mobile, achieving 5–10x gRPC performance with $<5 \mu\text{s}$ RDMA latency (50x lower than gRPC) and 1.7M op/s per core. Included FlowBuf, a zero-serialization library with Protocol Buffers-like APIs.
- Designed an RDMA orchestration framework to offload transactions of dependent communication verbs to distributed SmartNICs or Memory Processor.
- Enabled remote direct memory access without pinning memory, increasing effective memory size with tiered memory and accelerating application initialization.
- Designed new abstractions for remote memory access: connectionless, scalable and reliable Read, Write, Send and Recv operations, and flexible atomic operations.
- Designed a Memory-Level Cache utilizing local DDR as a cache of remote memory, improving Redis performance by 21%~45% and making Redis on remote memory only 2.5% slower than local.

AKG: Automatic Kernel Generation for NPUs using Polyhedral Transformations

HUAWEI

Published in PLDI '21 and TOCS '24; open sourced at <https://atomgit.com/mindspore/akg>; advised by Dr. Peng Di

June '19 – Apr. '20

- Architect and technical leader of the dynamic shape team (15 developers); one of the key architects of the polyhedral compilation framework.
- Built a tensor compiler for neural processing units (NPUs), replacing manually written schedule templates with ILP-based polyhedral schedulers to enable a wide class of loop transformations.
- Designed a novel composition of loop tiling and hierarchical fusion via schedule tree extensions, automating software-controlled data orchestration across the NPU's multi-level, multi-directional memory hierarchy and heterogeneous compute units.
- Achieved 1.6x speedup over TVM on single operators, 5.6x over hand-optimized C/C++ on fused subgraphs, while reducing development effort from months to hours.

SocksDirect: Datacenter Sockets can be Fast and Compatible

MICROSOFT RESEARCH ASIA

Published in SIGCOMM '19, first author; Global 1st place in IT Pros, Microsoft Hackathon; 2 US Patents granted; advised by Dr. Lintao Zhang

July '17 – May '19

- Designed a high-performance user-space socket system that replaces Linux sockets without changing any existing applications, solving the pervasive bottleneck of OS kernel consuming ~80% CPU time in datacenter network communication.
- Leveraged RDMA for inter-host and shared memory for intra-host communication. A trusted monitor daemon handles security and connection establishment while enabling peer-to-peer data plane communication between processes.
- Eliminated buffer copying, context switching, and thread synchronization, achieving 7–20x throughput, 17–35x latency improvement over Linux sockets, and 5.5x lower Nginx HTTP latency.
- Microsoft secured 2 US patents (US11,792,272 and US11,880,725) based on this technology.

KV-Direct: High-Performance Key-Value Store with Programmable NIC

MICROSOFT RESEARCH ASIA

Published in SOSP '17, first author; top 1% most cited in CS (2017); advised by Dr. Lintao Zhang

May '16 – Nov. '17

- Offloaded key-value store processing to programmable NICs, bypassing the CPU bottleneck. Extended RDMA primitives to key-value operations enabling remote direct access to host memory, with support for vector operations and user-defined functions.
- Achieved 180M ops/s per NIC (10x faster than CPU, 3x better energy efficiency) with $<10 \mu\text{s}$ latency. Built a 1.22 billion ops/s server with 10 FPGA NICs. Remains the state-of-the-art throughput record on a single server 8+ years after publication.
- Hid PCIe latency with optimized hash table, slab allocator, out-of-order execution, load dispatch and client-side batching.

ClickNP: Highly Flexible Network Processing with FPGA

MICROSOFT RESEARCH ASIA

Published in SIGCOMM '16, first author; top 1% most cited in CS (2016); advised by Dr. Kun Tan

July '15 – Oct. '16

- Developed the first FPGA-accelerated platform for general network functions, enabling developers to write network components in high-level language that is automatically translated into efficient hardware, resolving the critical trade-off between flexibility and performance.
- Achieved 40 Gbps line rate, 200M packets/s, and sub-2 μ s latency — 10x faster and 10x lower latency than software-only solutions. Enabled joint CPU-FPGA processing with 25 Gbps throughput and 1 μ s latency.
- Recognized as an official Microsoft Research project. The SmartNIC platform enabled a decade-long research program on RDMA scalability and datacenter acceleration, directly spawning 10+ follow-on publications.

Major Publications

1Pipe: Scalable Total Order Communication in Data Center Networks

Bojie Li, Gefei Zuo, Wei Bai and Lintao Zhang

Proceedings of the 2021 ACM SIGCOMM Conference (SIGCOMM'21)

AKG: Automatic Kernel Generation for Neural Processing Units using Polyhedral Transformations

Jie Zhao, **Bojie Li**, Wang Nie, Zhen Geng, Renwei Zhang, Xiong Gao, Bin Cheng, Chen Wu, Yun Cheng, Zheng Li, Peng Di, Kun Zhang and Xuefeng Jin

Proceedings of the 42nd ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI'21)

SocksDirect: Datacenter Sockets Can be Fast and Compatible

Bojie Li, Tianyi Cui, Zibo Wang, Wei Bai and Lintao Zhang

Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM'19)

KV-Direct: High-Performance In-Memory Key-Value Store with Programmable NIC

Bojie Li, Zhenyuan Ruan, Wencong Xiao, Yuanwei Lu, Yongqiang Xiong, Andrew Putnam, Enhong Chen and Lintao Zhang

Proceedings of the 26th ACM Symposium on Operating Systems Principles (SOSP'17)

ClickNP: Highly Flexible and High-performance Network Processing with Reconfigurable Hardware

Bojie Li, Kun Tan, Layong (Larry) Luo, Yanqing Peng, Renqian Luo, Ningyi Xu, Yongqiang Xiong, Peng Cheng and Enhong Chen

Proceedings of the 2016 ACM conference on SIGCOMM (SIGCOMM'16)

Other Publications (Selected)

Fast and Scalable Selective Retransmission for RDMA

Peng Huang, Guo Chen, Xin Zhang, Chengcheng Liu, Han Wang, Hao Shen, Yicheng Bian, Yuanwei Lu, Zhenyuan Ruan, **Bojie Li**, Binzhang Fu, Kun Tan

IEEE INFOCOM 2025 - IEEE Conference on Computer Communications (INFOCOM'25)

Advancing RDMA Scalability With High Performance

Xijin Yin, Guo Chen, Xizheng Wang, Bo Wang, Huichen Dai, **Bojie Li**, Binzhang Fu, Kun Tan

IEEE/ACM Transactions on Networking (ToN), 2025.

Modeling the Interplay between Loop Tiling and Fusion in Optimizing Compilers Using Affine Relations

Jie Zhao, Jinchun Xu, Peng Di, Wang Nie, Jiahui Hu, Yanzhi Yi, Sijia Yang, Zhen Geng, Renwei Zhang, **Bojie Li**, Zhiliang Gan, Xuefeng Jin

ACM Transactions on Computer Systems (TOCS), Volume 41, Issue 1-4, 2024.

FastWake: Revisiting Host Network Stack for Interrupt-mode RDMA

Bojie Li, Zihao Xiang, Xiaoliang Wang, Han Ruan, Jingbin Zhou, Kun Tan

Proceedings of the 7th Asia-Pacific Workshop on Networking (APNet'23)

StaR: Breaking the Scalability Limit for RDMA

Xizheng Wang, Guo Chen, Xijin Yin, Huichen Dai, **Bojie Li**, Binzhang Fu, Kun Tan

2021 IEEE 29th International Conference on Network Protocols (ICNP'21)

Towards Stateless RNIC for Data Center Networks

Pulin Pan, Guo Chen, Xizheng Wang, Huichen Dai, **Bojie Li**, Binzhang Fu, Kun Tan

Proceedings of the 3rd Asia-Pacific Workshop on Networking (APNet'19)

Dedas: Online Task Dispatching and Scheduling with Bandwidth Constraint in Edge Computing

Jiaying Meng, Haisheng Tan, Chao Xu, Wanli Cao, Liuyan Liu, **Bojie Li**

IEEE INFOCOM 2019 - IEEE Conference on Computer Communications (INFOCOM'19)

MP-RDMA: Enabling RDMA with Multi-Path Transport in Datacenter

Guo Chen, Yuanwei Lu, **Bojie Li**, Kun Tan, Yongqiang Xiong, Peng Cheng, Jiansong Zhang, Thomas Moscibroda

IEEE/ACM Transactions on Networking (ToN), Volume 27, Issue 6, 2019.

Online Deadline-Aware Task Dispatching and Scheduling in Edge Computing

Jiaying Meng, Haisheng Tan, Xiang-Yang Li, Zhenhua Han, **Bojie Li**

IEEE Transactions on Parallel and Distributed Systems (TPDS), Volume 31, Issue 6, 2019.

Joint Heterogeneous Server Placement and Application Configuration in Edge Computing

Jiaying Meng, Chaoting Zeng, Haisheng Tan, Zhenhua Li, **Bojie Li**, Xiang-Yang Li

2019 *IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS'19)*

FUSO: Fast Multi-Path Loss Recovery for Data Center Networks

Guo Chen, Yuanwei Lu, Yuan Meng, **Bojie Li**, Kun Tan, Dan Pei, Peng Cheng, Layong Luo, Yongqiang Xiong, Xiaoliang Wang, Youjian Zhao

IEEE/ACM Transactions on Networking (ToN), 2018

ST-Accel: A High-Level Programming Platform for Streaming Applications on FPGA

Zhenyuan Ruan, Tong He, **Bojie Li**, Peipei Zhou, Jason Cong

IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM'18)

Multi-Path Transport for RDMA in Datacenters

Yuanwei Lu, Guo Chen, **Bojie Li**, Kun Tan, Yongqiang Xiong, Peng Cheng, Jiansong Zhang, Enhong Chen and Thomas Moscibroda

Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI'18)

The Feniks FPGA Operating System for Cloud Computing

Jiansong Zhang, Yongqiang Xiong, Ningyi Xu, Ran Shu, **Bojie Li**, Peng Cheng, Guo Chen, Thomas Moscibroda

Proceedings of the 8th Asia-Pacific Workshop on Systems (APSys'17)

Memory Efficient Loss Recovery for Hardware-based Transport in Datacenter

Yuanwei Lu, Guo Chen, Zhenyuan Ruan, Wencong Xiao, **Bojie Li**, Jiansong Zhang, Yongqiang Xiong, Peng Cheng, Enhong Chen

Proceedings of the First Asia-Pacific Workshop on Networking (APNet'17)

Fast and Cautious: Leveraging Multi-path Diversity for Transport Loss Recovery in Data Centers

Guo Chen, Yuanwei Lu, Yuan Meng, **Bojie Li**, Kun Tan, Dan Pei, Peng Cheng, Layong (Larry) Luo, Yongqiang Xiong, Xiaoliang Wang, Youjian Zhao

Proceedings of the 2016 USENIX Annual Technical Conference (ATC'16)

Patents**Hardware-based Memory Compression**

Lintao Zhang, John G. Bennett, **Bojie Li**

US Patent US12430061B2, 2025.

Communication Method, Apparatus, and System, and Storage Medium

Hao Shen, Guo Chen, **Bojie Li**

US Patent US20250370862A1, 2025.

Request Processing Method, Device and System

Daoxin Li, Xin Zhang, **Bojie Li**

EP Patent EP4535149A4, 2025.

Memory Access Method and Related Device

Bojie Li, Adi Geron, Sagi Rabinovitch

US Patent US20250085853A1, 2025.

Establishment of Queue between Threads in User Space

Bojie Li, Tianyi Cui, Zibo Wang, Wei Bai, Lintao Zhang

US Patent 11,880,725, 2024.

Network Device-Based Data Processing Method and Network Device

Bojie Li, Lei Zhou

US Patent US20240427598A1, 2024.

Establishment of Socket Connection in User Space

Bojie Li, Tianyi Cui, Zibo Wang, Wei Bai, Lintao Zhang

US Patent US11792272B2, 2023.

Implementation of Sockets in User Space

Bojie Li, Tianyi Cui, Zibo Wang, Wei Bai, Lintao Zhang

WO Patent WO2020096870A1, 2020.

Total-Order Message Mechanism in a Distributed System

Lintao Zhang, Wei Bai, Gefei Zuo, **Bojie Li**

WO Patent WO2019226367A1, 2019.

Professional Activities

Program Committee member, APNet 2026
Program Committee member, ChinaSys 2025
Program Committee member, APNet 2025
Program Committee member, APSys 2024

Awards in USTC & MSRA

ACM China Doctoral Dissertation Award (2 recipients nationwide) *Dec. '19*
Microsoft Research Asia Fellowship (10 recipients across Asia) *Oct. '17*
Chinese Academy of Sciences Presidential Scholarship (中科院院长奖学金) (400 recipients nationwide)
June '19
Microsoft Hackathon 2017 - Global First Place in IT Pros Category (1 of 200+ participants) *Aug. '17*
Microsoft Hackathon 2016 - Global Second Place in Cloud & Enterprise (2 of 1000+ participants) *Aug. '16*
Microsoft Hackathon 2017 - Most Impactful Project Award, Beijing (1 of 200+ participants) *Aug. '17*
MSRA Student Techfest 2016 - Best Presentation Award (1 of 50+ presenters) *Oct. '16*

Awards before Undergraduate Study

National Olympiad in Mathematics, Hebei Province 2009 - First Prize (Top 10) *Oct. '09*
National Olympiad in Informatics (NOI) 2009 - National Bronze Medal *Jan. '09*
National Olympiad in Informatics, Hebei Province (NOIP) 2008 - First Prize (Ranked 2nd) *Nov. '08*
Hua Luogeng Cup National Mathematics Competition - Gold Medal (National Top 10) (第九届华罗庚金杯少年数学邀请赛全国总决赛金牌) *May '03*