BAODI SHAN

Stony Brook, NY
+1 (631) 559-0063 ◇ baodi.shan@stonybrook.edu ◇ GitHub: lwshanbd

SUMMARY

Ph.D. candidate in Computer Science focused on networking and system architecture for large-scale AI and HPC systems. Research spans GPU-initiated communication, RDMA and PGAS runtime design, and cross-layer performance analysis for distributed GPU clusters over InfiniBand and Slingshot. Strong publication record and hands-on prototyping experience in C++ and Python, with demonstrated improvements in communication latency, scalability, and end-to-end application efficiency.

EDUCATION

Stony Brook University Aug 2021 – Sep 2026 (expected)
Ph.D. in Computer Science *Stony Brook, NY*

Shandong University Sep 2017 – Jun 2021
B.Eng. in Computer Science and Technology *Qingdao, China*

SKILLS

Research Areas	Networking and system architecture, distributed GPU systems, communication runtimes
Interconnects	RDMA (InfiniBand/RoCE), Slingshot, libibverbs, libfabric, UCX
Runtime and GPU	GASNet-EX, NCCL, NVSHMEM, CUDA, HIP, OpenMP Target
Programming	C, C++17/20, Python
Systems and Tools	Linux, LLVM/Clang, CMake, Slurm, Flux, profiling, tracing

EXPERIENCE

Stony Brook University Aug 2021 – Present
Research Assistant *Stony Brook, NY*

- Architected GICC, a GPU-initiated communication and coordination runtime spanning libfabric-based Slingshot and libibverbs-based InfiniBand backends, enabling GPU-triggered progress, active messages, and barrier operations across heterogeneous GPU clusters.
- Co-designed transport and synchronization mechanisms for distributed GPU communication, achieving up to 1.95x lower put latency than NVSHMEM on InfiniBand and up to 25% higher weak-scaling efficiency on Slingshot.
- Demonstrated end-to-end gains on a 64-GPU stencil workload, reducing communication time by over 52% versus GPU-aware MPI and improving parallel efficiency from 35.4% to 42.0%.
- Built a PGAS-based distributed GPU runtime over GASNet-EX and redesigned RDMA data paths, increasing bandwidth by 25% and reducing end-to-end latency by 45%.
- Refactored remote OpenMP offloading runtimes over MPI and UCX, removing host synchronization bottlenecks and delivering up to 70% speedup on distributed heterogeneous GPU workloads.

Amazon Web Services May 2025 – Aug 2025
Summer Intern *New York, NY*

- Built an evaluation-driven multi-agent system for automated security assessment using MCP, AWS Lambda, and Fargate, including private retrieval, vector indexing, and benchmarking pipelines that reached 0.92 F1 on internal tasks.

Lawrence Livermore National Laboratory May 2024 – Aug 2024
Summer Intern *Remote*

· Designed Fuzzlang, a Clang-based framework that combines systematic compiler-error generation with an LLM agent, improving code-correction accuracy after fine-tuning from 37.22% to 93.97% on Llama3-8B and from 72.29% to 96.70% on GPT-4o-mini; contributed to ComPile V2 C/C++, a large-scale LLVM-based IR dataset for compiler research.

TotalEnergies May 2023 – Aug 2023
Summer Intern *Houston, TX*

· Optimized stencil and sparse linear algebra workloads for heterogeneous HPC systems, improving computational efficiency and deploying scalable solvers on production PANGEA II and PANGEA III clusters.

PUBLICATIONS

Baodi Shan, Mauricio Araya-Polo, and Barbara Chapman. "GICC: A High-Performance Runtime for GPU-Initiated Communication and Coordination in Modern HPC Systems." HPDC 2026.

Baodi Shan, Mauricio Araya-Polo, and Barbara Chapman. "DiOMP-Offloading: Toward Portable Distributed Heterogeneous OpenMP." SC-W 2025.

Baodi Shan, Mauricio Araya-Polo, Johannes Doerfert, and Barbara Chapman. "Discussion of Device-Device Collective Communication in OpenMP Target Offloading." IWOMP 2025.

Baodi Shan, Mauricio Araya-Polo, and Barbara Chapman. "Towards a Scalable and Efficient PGAS-based Distributed OpenMP." IWOMP 2024.

Baodi Shan, Mauricio Araya-Polo, and Barbara Chapman. "Evaluation of Directive-based Programming Models for Stencil Computation on Current GPGPU Architectures." IWOMP 2024.

Baodi Shan, Mauricio Araya-Polo. "Evaluation of Programming Models and Performance for Stencil Computation on GPGPUs." IPDPS 2024.

Baodi Shan, Mauricio Araya-Polo, Abid M. Malik, and Barbara Chapman. "MPI-based Remote OpenMP Offloading: A More Efficient and Easy-to-use Implementation." PPoPP-W 2023.

Wenbin Lu, **Baodi Shan**, Eric Raut, Jie Meng, Mauricio Araya-Polo, Johannes Doerfert, Abid M. Malik, and Barbara Chapman. "Towards Efficient Remote OpenMP Offloading." IWOMP 2022.

Baodi Shan, Aabid Shamji, Jiannan Tian, Guanpeng Li, and Dingwen Tao. "LCFI: A Fault Injector for Studying Lossy Compression Error Propagation in HPC Programs." IEEE Big Data Workshops 2020.

PROFESSIONAL SERVICE

OpenMP Architecture Review Board: Stony Brook representative (2024 – Present).

Argonne Leadership Computing Facility Director's Discretionary Program: PI (2025 – Present).

Artifact Evaluation Program Committee: CGO 2025, MobiSys 2024, MobiSys 2025.

Session Chair: PMAM 2024; Expanding Horizons in AI with HPC Workshop 2025.