LONG CHENG

Senior Architect — LLM Inference Infrastructure — GPU Kernels — System-Algorithm Co-Design
China — buaalongcheng@gmail.com — +86 13126525737
chenglong92.github.io

Professional Summary

Senior architect specializing in LLM inference infrastructure, GPU kernel optimization, and system-algorithm co-design for frontier AI systems. Currently building TensorRT-LLM infrastructure for NVIDIA's latest GPUs, spanning sparse attention, lossless Top-K, and LLM for system and algorithm design and optimization, from GPU kernels to compiler workflows such as LLM4Compiler. Previously led low-precision LLM operator and LLM-compiler initiatives at Huawei, delivering measurable gains across AI and HPC workloads. PhD-trained in numerical methods and high-performance computing, with a track record of translating research into production systems, open-source software, and measurable performance wins.

Core Expertise

TensorRT-LLM, CUDA/C++, PyTorch, Triton, GPU kernel optimization, sparse attention, long-context LLM inference, mixed precision (FP8/FP16/NVFP4), LLM-native algorithm design, AI-assisted system and algorithm optimization, compiler optimization, LLM4Compiler, auto-vectorization, ARM NEON/SVE, HPC, numerical linear algebra

Professional Experience

NVIDIA, China Nov 2025 – Present
Senior Architect, Compute DL Architecture

- Build TensorRT-LLM inference infrastructure for the latest NVIDIA GPUs, improving long-context LLM serving through system-algorithm co-design.
- Designed and upstreamed the GVR (Guess-Verify-Refine) Top-K algorithm and kernel for DSA sparse attention in TensorRT-LLM: a four-stage, lossless, search-based, sort-free, almost atomic-free design for long-context inference.
- Delivered 1.88x average kernel-level speedup on real DeepSeek-V3.2 decode workloads on NVIDIA Blackwell GPUs, with up to 2.42x per-layer speedup.
- Improved end-to-end generation throughput by up to 9.3% on the low-latency region of the long-context SWE-bench Pareto frontier (DeepSeek-V3.2, ISL/OSL=101K/1K, GB200); high-throughput configurations currently reuse the baseline radix-based Top-K path pending further scheduling optimization.
- Developed the kernel through AI-assisted, human-in-the-loop co-design and added production-ready metadata, scratch-buffer, and multi-row / MTP decode support for TensorRT-LLM integration.
- Merged the feature into NVIDIA/TensorRT-LLM mainline through PR #12385 and presented the work in a public NVIDIA TensorRT-LLM technical blog.

Huawei Beijing Research Center, Beijing, China Sep 2023 – Nov 2025
Principal Research Engineer, Compilers and Programming Languages Lab

- Led research on high-performance LLM operators, mixed-precision algorithms, and LLM-guided compiler optimization.
- Led PASA (Pseudo-average Shifting Attention), a robust low-precision attention algorithm for LLM inference on Ascend NPUs, contributing up to 1.65x end-to-end acceleration in customer-facing workloads.
- Led the open-source LLM4Compiler initiative under openEuler AI4C, including VecTrans and G2CTrans, to apply LLMs to system, algorithm, and compiler optimization workflows.
- Developed VecTrans, an iterative-refinement LLM transformation framework for auto-vectorization on ARM NEON/SVE CPUs, surpassing the prior state of the art on TSVC-2 and resulting in an open-source release plus a CGO 2026 submission.
- Explored CPU-NPU heterogeneous acceleration for mobile-game physics workloads through mathematically equivalent transformations and tensorized compilation.
- Collaborated with Huawei overseas teams and academic partners on long-term LLM compiler and sparse tensor optimization projects.

Huawei Beijing Research Center, Beijing, China Apr 2021 – Aug 2023
Researcher, Heterogeneous Programming Group

- Developed mixed-precision dense LU decomposition across Ascend AI processors and Kunpeng CPUs, delivering more than 18x speedup for boundary-element workloads; the work was showcased at Huawei Connect 2021.
- Developed a massively parallel single-double mixed-precision sparse solver for high-order finite-element Navier-Stokes simulations, achieving more than 3x speedup over double-precision baselines.
- Worked on mixed-precision algorithm design at the intersection of numerical linear algebra, scientific computing, and heterogeneous hardware acceleration.

Huawei Cambridge Research Center / HiSilicon, Cambridge, UK Jan 2020 – Apr 2020
PhD Intern Researcher

- Researched scientific computing and mixed-precision methods for PDE workloads in collaboration with UK-based teams.

Selected Open-Source, Publications, and IP

- NVIDIA/TensorRT-LLM PR #12385, "GVR (Guess-Verify-Refine) / Temporally-Correlated Heuristic-guided Indexer TopK for Sparse Attention," merged into main in Apr 2026.
- openEuler AI4C / LLM4Compiler: led the open-source project, including VecTrans and G2CTrans, for LLM-driven system, algorithm, and compiler optimization workflows.
- NVIDIA TensorRT-LLM Technical Blog (2026), "Temporal Correlation Meets Sparse Attention: A Heuristic Top-K Kernel for Blackwell."
- Zheng, Cheng*, et al. (2025). "VecTrans: LLM Transformation Framework for Better Auto-vectorization on High-Performance CPUs." arXiv:2503.19449, CGO 2026 submission.
- Cheng*, Liao, et al. (2025). "Online Pseudo-average Shifting Attention (PASA) for Robust Low-precision LLM Inference: Algorithms, Numerical Analysis and Performance." arXiv:2503.01873, NeurIPS 2025 submission.
- Huawei High-value Patent 2025: "Method for Determining Model Output Results and Related Devices."

Education

Beihang University (BUAA), Beijing, China 2014 – 2021
PhD, Aerospace Engineering — Supervisor: Prof. Xiaofeng Sun

- Supported by the National Natural Science Foundation of China; developed a massively parallel CFD/CAA solver for turbomachinery flow-sound interaction.

University of Cambridge, Cambridge, UK 2019 – 2020
CSC-Sponsored Full-time Visiting PhD Student, Computational and Applied Mathematics — Host Supervisor: Prof. Paul Tucker

- Worked with Prof. Paul Tucker on high-performance scientific computing workloads on UK supercomputing systems.

Northwestern Polytechnical University (NWPU), Xi'an, China 2010 – 2014
BEng, Aerospace Engineering

Selected Awards

- Silver Award for Innovation Pioneer, Huawei Central Software Institute (2024)
- Huawei Golden Medal Team Award, Huawei's highest team honor (2024)
- Software Elite Award, Huawei (2023)
- Best New Employee, Compilers and Programming Languages Lab, Huawei (2022)
- Meritorious Winner, Mathematical Contest in Modeling (MCM), SIAM & COMAP (2013)
- National Scholarship, Ministry of Education of China (top 2.2%, 2013)