Sreeram Vennam

+1 (412) 670-2975 | Pittsburgh, PA
svennam@andrew.cmu.edu

linkedin.com/in/**vnnm**
github.com/**vnnm404**
scholar.google.com/**vnnm**

Education

Carnegie Mellon University Pittsburgh, PA
Master of Science in Computer Science (MSCS) Dec 2026

- CGPA: **4.09/4.0**.
- Coursework: Compiler Design (15-611), Parallel Programming (15-618), Graphical Models (10-708), Advanced Machine Learning Systems (15-779), Distributed Systems (15-640), Database Systems (15-645).
- TA: LLM Systems (11-868)

International Institute of Information Technology - Hyderabad Hyderabad, India
Bachelor of Technology in Computer Science and Engineering (Honors) May 2025

- CGPA: **9.35/10.0**.

Skills

Programming Languages: Python, C/C++, Go, Java
Technologies: PyTorch, CUDA, Numpy, Pandas, Einops, Accelerate, vLLM, BitsAndBytes, PyG, Linux, Git, Bash, LaTeX

Experience

Google Hyderabad, India
Software Engineering Intern May 2024 - Aug 2024

- Designed and built a scalable connector for migrating large datasets from SQLServer to BigQuery (GCP), and built a novel partitioning strategy increasing data transfer speeds by **15%** specific to SQLServer.
- Optimized reliability by adding unit tests (**80% coverage**), end-to-end integration tests, and pipeline health probers.
- Resolved a critical bug preventing potential runtime crashes in production, enhancing system reliability across **10+** connectors. Repaired failing integration tests (including FacebookAds), improving CI pipeline stability and increasing test coverage by **22%**.

Selected Publications

LLM Vocabulary Compression for Low-Compute Environments Neural Compression, NeurIPS 2024
Sreeram Vennam, Anish Joishy, Ponnurangam Kumaraguru

- Reduced the embedding layer overhead in SLMs, achieving 3× throughput and 3.4× lower peak memory.

Higher Order Structures For Graph Explanations AAAI 2025
Akshit Sinha*, **Sreeram Vennam***, Charu Sharma, Ponnurangam Kumaraguru

Projects

gpt2.cu GPT-2 training in a single megakernel. vnnm404/gpt2.cu

- First to implement a **working Megakernel for GPT-2 training** to improve training throughput.
- Built all forward, backward, and AdamW update kernels in custom CUDA (no vendor libraries); optimized matmul and layernorm using shared memory, warp tiling, vectorized loads, double buffering, and warp-level reductions, achieving performance competitive with or exceeding cuDNN and reducing training step time from **500 ms to 170 ms**.

mini-flash-attention Flash attention implemented in CUDA. vnnm404/mini-flash-attn

- Implements a minimal version of Flash Attention in CUDA as a C/C++ extension to PyTorch. Demonstrates a **9x** speedup on GPU for a sequence length s=1024 and hidden dimension size of d=32 over standard PyTorch implementations.

bustub A multi version relational database management system.

- Built core database internals including a concurrent buffer pool manager, B+-Tree indexes, query optimizer, and optimistic MVCC.
- Achieved **1st** place on the Query Execution leaderboard by optimizing joins and aggregations via column pruning, join reordering, hot-path execution, and Bloom-filter based sideways information passing (SIP).
- Implemented the Grace Hash Join which enabled **sub 1 second** (684 ms) execution times for a 3 way join involving **10 million rows** per table.

gradf Reverse mode automatic differentiation in C. vnnm404/gradf

- Implements reverse mode automatic differentiation in C with a simple printf-like API.