SWEBERT CORREA

linkedin · correaswebert@gmail.com · (470) 685-1680 · github

Education

Georgia Institute of Technology, Atlanta — Aug 2024 - May 2026
Master of Science in Computer Science | Graduate Teaching Assistant — GPA: 4.0/4.0

College of Engineering, Pune — Aug 2018 - May 2022
Bachelor of Technology in Computer Engineering | Minor in Financial Engineering — GPA: 9.05/10.0

Coursework: Advanced Operating Systems, GPU HW-SW, Database Implementation, Distributed Computing

Skills

Languages: C, C++, Python, CUDA, Rust, Go (Golang), x86 Assembly, Bash, SQL

Infrastructure: Kubernetes, Docker, AWS (S3, EC2, Lambda, SQS), PostgreSQL, DynamoDB, Git, Jenkins

Systems: Valgrind, GDB, Strace, NVIDIA Nsight, Flame Graph

Experience

AMAZON — Seattle, WA
Software Development Engineer Intern | Kotlin, Fargate, DynamoDB, SQS — May 2025 - Aug 2025

- Engineered a real-time AI-guided device setup notification service for Alexa+ handling **millions of DAU**
- Integrated multimodal **adaptive push/pull delivery** boosting render rate from 3.84% to over 50%
- Saved 2 weeks (5× reduction in time) for each on-boarded service to send push, in-app and voice notifications

RAKUTEN (ROBIN.IO) — Pune, India
Software Engineer | C, Python, Kubernetes, PostgreSQL — Jun 2022 - Jul 2024

- Engineered high-performance in-house object storage solution **tailored for AI workloads** to replace MinIO, achieving **$4M+ annual cost savings** and reducing operational overhead across global clusters.
- Architected a **disk rebuild subsystem** capable of recovering data across **4PB of storage** by implementing an erasure coding subsystem using `liberasurecode` and disk-sets to parallelize recovery across 250+ drives.
- Developed a high-availability RPC-based **cluster manager over Kubernetes** for workload-based microservice lifecycle management such as data servers, metadata servers and gateways.
- Developed an asynchronous API gateway with distributed caching and a custom HTTP **byte-stream processor** using GNU `libmicrohttpd` to handle real-time, **NUMA-aware erasure coding**.
- Spearheaded the resumable multipart upload for parallelized data transfer and content-addressable chunk **deduplication engine** to optimize storage utilization and reduce I/O overhead.
- Mentored an intern to integrate high-resolution telemetry in a **Grafana** dashboard for cluster health.

Projects

Xen Hypervisor Credit Scheduler | Operating Systems, C, Linux — Georgia Tech, Fall '25

- Implemented type-2 hosted hypervisor leveraging Linux signals for context switching and thread-local storage
- Built M:N user-level threading library with a credit-based scheduler supporting load balancing across vCPUs

LLM Inference Optimization on GPU | GPU Architecture, CUDA, PyTorch — Georgia Tech, Fall '25

- Developed a custom **FlashAttention** kernel in CUDA with kernel fusion to minimize off-chip memory traffic.
- Designed a tiling mechanism using shared memory and online softmax for numerical stability and. scalability
- Optimized **KV cache** management to reduce TBT, TTFT & latency and the increase context window.

BuzzDB Database Implementation | Database Internals, C++, Valgrind — Georgia Tech, Spring '25

- Implemented a WAL in an ARIES-compliant log manager along with fuzzy checkpointing optimization.
- Built a lock manager supporting 2-Phase Locking protocol to enable high-concurrency transaction processing.
- Designed a query optimizer that utilizes catalog statistics, selectivity estimation, and join-ordering.