Thomas Braun

Senior Rust Systems Engineer | LLM Inference, GPU Kernels, Post-Quantum Peer-to-Peer

Charlotte, North Carolina, US | thomas@avarok.net | thomaspbraun.com | github.com/tbraun96 | linkedin.com/in/tpbraun | blog.thomaspbraun.com

Summary

Rust systems engineer and founder with eight years building a family of inter-related open source products: the Atlas Inference Engine (pure Rust and CUDA LLM inference, submitted to MLPerf Inference v6.1 on NVIDIA and AMD), the patented post-quantum Citadel Protocol, the Citadel Workspace built on it, and the browser-side tools around them. Expertise in GPU kernel optimization, quantization, applied post-quantum and threshold cryptography, asynchronous networking and peer-to-peer systems.

Experience

Co-Founder | Atlas Cybernetics Corp, Atlas Inference Engine | Mar 2026 - Present

- Created an open-source LLM inference engine in pure Rust and CUDA that serves OpenAI-, Anthropic- and Responses-compatible APIs from a single ~75 MB binary with no Python, PyTorch or runtime compilation.
- Hand-tuned CUDA kernels per hardware, model and quantization target (NVFP4 and FP8 quantization, MTP speculative decoding, radix-tree prefix caching, expert parallelism across nodes) for NVIDIA GB10, compiled unchanged for AMD Strix Halo through SCALE.
- Out-served vLLM at all 8 concurrency levels measured (1 to 128) on DGX Spark; 2-3x faster than vLLM on the Qwen 3.5 and 3.6 35B-A3B and 27B models.

Founder & Rust Consultant | Avarok Cybersecurity, Remote | Jan 2018 - Present

- Designed and implemented The Citadel Protocol, a post-quantum peer-to-peer and client-server networking SDK in 100% safe Rust on Tokio, with built-in NAT traversal, TCP/TLS/QUIC transports and a Remote Encrypted Virtual Filesystem.
- Invented and implemented the patented Recursive Cryptography Protocol (US 12,224,993): multi-layered encryption with 3D-matrix ratcheting, per-message re-keying and Kyber key exchange for crypto-agility and true perfect forward secrecy.

Principal Rust Software Engineer | Nomyx, Remote | Feb 2025 - Mar 2026

- Implemented ERC-2535 (the Diamond Proxy Standard) for the Stellar network, giving granular control over smart-contract upgradeability with shared storage so contracts governed by one proxy communicate through memory.
- Authored Stellar Ecosystem Proposals (SEPs) and upstream pull requests for shared-storage integration between diamond facets.

Senior Software Engineer | Synopsys, formerly Ansys, Remote | Nov 2021 - Mar 2026

- Designed and solely implemented Synopsys Data Shield, a CP-ABE based proxy and vector/graph database encryptor that keeps company IP encrypted at rest and decrypted only for use in agentic RAG.

Rust Software Engineer | Commonwealth Labs, Remote | Dec 2022 - Feb 2025

- Implemented a threshold distributed key generator (DKG) in Rust using multi-party ECDSA and WT-FROST for voting and signing proposals on a distributed system.

Software Engineer I | Kraken Digital Asset Exchange, Core Backend - Security, Remote | Oct 2022 - Dec 2022

- Implemented RPC microservice endpoints in Rust for two-factor authentication and device approval.

Rust Software Engineer | Webb.tools, Remote | Nov 2021 - Oct 2022

- Converted a rounds-based serial protocol into a parallel, asynchronous protocol, doubling efficiency and performance.

Solidity Developer | Bitgem NFT, Remote | Dec 2020 - Feb 2021

- Developed a lottery algorithm using dynamic-difficulty bitmasks that lets transactions randomly award bonus tokens.

Projects

Atlas Inference Engine | Founder and lead engineer | atlasinference.io | Mar 2026 - Present

- Pure Rust and CUDA: a single ~75 MB binary, no Python or PyTorch runtime, no runtime compilation
- Hand-tuned CUDA kernels per hardware, model and quantization target; ~170 kernel sources for GB10, compiled unchanged for AMD Strix Halo through SCALE
- 200K+ monthly unique visitors (atlasinference.io); 737 GitHub stars.

The Citadel Protocol | Creator and maintainer | avarok.net | Jan 2022 - Present

- Post-quantum key exchange with Kyber (default) and NTRU; AES-256-GCM, ChaCha20-Poly1305 and Ascon-80pq ciphers
- 131,510 crates.io downloads across the 9 published crates; 166 GitHub stars.

Citadel Workspace | Founder and lead engineer | work.avarok.net | May 2023 - Present

- End-to-end encrypted direct and group messaging with offline delivery, voice and video calls, and direct file transfer between peers
- 1,496 commits on the main repository; 30 releases, latest agent-v0.13.1.

LatticeDB | Creator and maintainer | avarok-cybersecurity.github.io/lattice-db | Jan 2026 - Present

- HNSW index with SIMD distance kernels, product and scalar quantization; BFS/DFS traversal, Cypher queries and weighted edges
- 2x faster vector search than Qdrant, both over HTTP (168 us against 330 us); 13x faster Cypher MATCH than Neo4j, HTTP against Bolt (85 us against 1,147 us).

CertifiedCopy | Creator | certified.sh

- 100% client-side: the parser, the SQLite engine and the PDF renderer are compiled to WebAssembly and run on the reader's machine; the test suite asserts it on every change

synchronized.work | Creator | synchronized.work | Nov 2026 - Present

- Agents join the tree through an MCP server; a remote prompt reaches an agent only if it is signed with ML-DSA-65

GeoHarness | Creator | geo.avarok.net | Nov 2026 - Present

- 29 evidence-graded checks in 5 scoring categories (AI access, indexability, markdown delivery, extractability, signals), each cited to a published source

heic-rs | Author | github.com/tbraun96/heic-rs | Sep 2026 - Present

- No libheif, no C toolchain, no unsafe
- 4.28x faster than the AGPL heic crate on a 2048x1536 photo (1.77 ms against 7.58 ms); 283,637 crates.io downloads.

Skills

Languages: Rust, CUDA, Python, TypeScript, JavaScript, Java, Solidity, SQL

AI Inference & GPU: LLM inference engines, CUDA kernel optimization, NVFP4 and FP8 quantization, Speculative decoding (MTP), Expert parallelism, Prefix caching, vLLM, MLPerf Inference, NVIDIA DGX Spark (GB10), AMD Strix Halo

Cryptography & Security: Post-quantum cryptography (ML-KEM/Kyber, NTRU, ML-DSA), Threshold cryptography (DKG, FROST), Ratcheting and perfect forward secrecy, AES-256-GCM, ChaCha20-Poly1305, CP-ABE, WebAuthn, NIST PQC standards

Networking & Distributed Systems: Async Rust (Tokio), Multithreading, TCP/UDP, QUIC, TLS, NAT traversal (STUN/TURN), Peer-to-peer architecture, WebRTC and WebCodecs, CRDTs, RPC microservices

Web, WebAssembly & Cloud: WebAssembly, React, Dioxus, Progressive Web Apps, Cloudflare Workers, Docker, Kubernetes, GitHub Actions, Playwright, Lighthouse

Education

Bachelor of Science, Geospatial Sciences and Geography | Oregon State University | 2020

Publications

Recursive Cryptography Protocol | US Patent 12,224,993 B2, Avarok Corp | Feb 2025

- A post-quantum session-layer protocol with highly configurable security: per-message re-keying, multi-layered encryption, 3D-matrix ratcheting and post-quantum key exchange inside a networking protocol for peer-to-peer messaging and network applications.

Awards

Delaware Patent of the Month | Swanson Reed

- Recognized as Patent of the Month by Swanson Reed, the largest R&D tax advisory firm in the United States.