

Bartek Ogryczak

pronounced bar-tech og-ree-chuck

Distributed Systems & AI Infrastructure Engineer

📍 Concord, CA ✉ b.ogryczak@gmail.com 💻 [bogryczak](#) 🗣 [vartec](#)

SUMMARY

Seasoned infrastructure engineer with 20+ years of programming experience, specializing in building and scaling high-traffic distributed services. Spent two years building and operating Claude inference infrastructure at massive scale at Anthropic, and now consults on applied large-language-model systems. Deep expertise in AWS, Python, and Rust. Proven track record in leading significant infrastructure projects and driving the adoption of best practices within technology teams.

SKILLS

CLOUD TECHNOLOGIES	AWS, GCP, Docker & Kubernetes
PROGRAMMING	Python, Rust, Django, Ruby
DATABASES	MySQL, Redis, PostgreSQL, BigQuery, ClickHouse
ARCHITECTURES	Microservices, Distributed Systems
AI INFRASTRUCTURE	LLMOps, Google Vertex, Google TPUs, Local & Self-Hosted Inference
OTHER	Kafka, Circuit Breakers, Rate Limiters, Protocol Buffers, Linux

EXPERIENCE

Technical Consultant, Stealth (AI × Legal) — San Francisco, CA

JUN 2026 – PRESENT

- Advising an early-stage venture on applying large language models in the legal domain.
- Architecture built around local and self-hosted inference, keeping confidential material inside the client's own environment.
- System design across model selection, evaluation, and data handling.

Details are limited while the company is in stealth.

Member of Technical Staff, Anthropic — San Francisco, CA

AUG 2024 – MAY 2026

- **Cloud Inference Team:** developed, operated, and scaled Claude inference infrastructure on Google Vertex AI. Scaled from near zero to one of Anthropic's largest enterprise inference channels.
- DRI for 6+ Claude model launches on Vertex AI.
- DRI for bringing Claude capabilities — including Thinking, Prompt Caching, Token Counting, and Multimodal Inputs — to Google Vertex AI.
- Core contributor to Dystro, Anthropic's inference coordination layer responsible for routing traffic across thousands of accelerators supporting large-scale Claude inference.
- Implemented key components of Vertex's Provisioned Throughput feature for Claude models.

Senior Software Engineer II, Sentry — San Francisco, CA

OCT 2023 – AUG 2024

- **Errors and Issues Team:** a cross-functional team responsible for building and maintaining Sentry's core product.
- Automated enforcement of logging best practices, preventing outages due to misconfiguration.
- Improved error grouping accuracy and maintainability within Sentry's core error-processing pipeline.
- Implemented caching to reduce ClickHouse data scans by up to 97%.
- Partnered with ML engineers to prototype LLM-based approaches for automated error analysis and triage.

Senior Software Engineer II, Chime — San Francisco, CA

2020 – 2023

- **Application Foundry Team:** developed tools to enhance service resilience and scalability. Managed both application code and infrastructure setup.
- AWS Migration: key role in a cross-functional “tiger” team that migrated infrastructure company-wide from RackSpace to AWS EKS.
- Standardized task scheduling by migrating to Kubernetes CronJobs.
- Developed Hawker, an in-house messaging system, improving cross-service communication.
- Implemented Twirp for standardized RPC, enhancing team interoperability.

SWE → Senior SWE II, Eventbrite — San Francisco, CA

2014 – 2020

- **Back-end Foundry Squad / Infrastructure Team:** led architectural solutions and system implementations to enable scalable operations.
- Designed and implemented Cerberus, an advanced permissions system.
- Co-created PySOA, enabling scalable service-oriented architecture and reducing remote call latency.

Back-end Developer, Layar — Amsterdam, The Netherlands

2011 – 2014

Designed and implemented scalable architecture for Layar's augmented reality applications.

Back-end Engineer / Software Analyst, Earlier Experience

1997 – 2011

- **Hyves (Amsterdam)** — back-end engineer at the Netherlands' largest social network; managed 2 billion image/video files with a replication algorithm that reduced latency and increased reliability.
- **Red Trabajar (Madrid)** — led a rewrite of Spain's largest job portal (Zend Framework, Oracle); built a system sending 15 million customized emails nightly.
- **Santander Bank (Madrid)** — real-time data integration for investment-banking risk management; devised an algorithm, running on a computing cluster, that cut latency and runtime 12-fold.
- **Melog.com (Warsaw)** — Linux sysadmin and back-end specialist on Oferty.net, a real-estate meta-search platform.

PATENTS

- Co-inventor (as Bartłomiej Ogryczak), [Staggered payload relay for pipelining digital payloads across network services](#) — US Patent 12,574,334 B2 (Chime Financial).

EDUCATION

University of Warsaw, MS in Computer Science — Warsaw, Poland

1995 – 2001

Specialized in Distributed Systems; developed USOSweb for the thesis — a component of the award-winning [USOS](#) system now used by over 50 universities.

LANGUAGES

Polish (native), English (fluent), Spanish (fluent)