

Josh Berns

Software engineer · Founder of Polynode
(acquired by [agg.market](#)) and MASHBOX Labs

Based in Bangkok, open to relocating
US citizen, no visa sponsorship needed

boshjerns@gmail.com
boshjerns.com
linkedin.com/in/josh-berns-6a553717b
github.com/boshjerns
x.com/boshjerns

SUMMARY

I fight the speed of light with software: real-time data infrastructure, on-device AI, native iOS apps and AI-built game worlds. I founded Polynode, a real-time Polymarket data API that [agg.market](#) acquired in July 2026, and built [agg.market](#)'s native app for iPhone, iPad and Mac. I also founded MASHBOX Labs. Self-taught, I went from music photography to smart contracts to shipping products. I contribute regularly to [llama.cpp](#) and work every day with parallel AI coding agents (Claude Code, Codex) behind review and verification gates across production systems.

EXPERIENCE

MASHBOX Labs · Founder

2026 to present

- Founded MASHBOX Labs to build MASHBOX: one engine that loads the worlds and play styles of games a player owns and mixes them into one game, with multiplayer.
- Released the Windows app in free early access; MASHBOX's own worlds run in the browser. mashbox.gg
- MASHBOX Labs is a member of NVIDIA Inception.

agg.market · Engineer

2026 to present

- Joined with the acquisition of Polynode and kept working across the product.
- Built the native [agg.market](#) app in Swift and SwiftUI for iPhone, iPad and Mac from one codebase, now in public beta on TestFlight.
- The app puts every prediction market venue on one screen, with smart order routing that splits each trade for the best price, and live crypto rounds.

Polynode · Founder and engineer

Oct 2025 to Jul 2026 · acquired by [agg.market](#)

- Built Polynode, a real-time Polymarket data API that decodes orders 3 to 5 seconds before they land onchain and streams them over WebSocket and REST.
- Built a historical archive of 1.2B+ fills across 2.7M wallets and 1M+ markets, queryable through a SQL explorer.
- Data layer for more than \$2B of Polymarket trading volume. Customers include [betmoar](#), [Gate](#), [PolyHelper.io](#), [POTS](#), [TerminalUp](#), [standtrade](#), [PolymarketScan.org](#), [Polygun](#), [Polycool](#) and [Share](#).

Mithril · Creator

Sep 2025 to present

- Built a free iPhone and iPad app that runs LLMs entirely on-device: Apple's on-device model plus open GGUF models from [Hugging Face](#). Rated 4.5 stars on the [App Store](#).
- Added memory-fit checks before each download, a live engine dashboard (tokens per second, time to first token, cache) and on-device speech with NVIDIA Parakeet.
- Gave it agent tools that act with permission through Reminders, Calendar, Messages, Maps and Shortcuts, and a live map of every network request.

Independent products and freelance

2022 to 2025

- Shipped AI products and sold two of them: [DoFDGPT](#), which turns government contract opportunities into briefs, and [SnapCycle](#), an AI recycling assistant.
- Quantish let AI agents trade prediction markets and moved about \$500K in volume.

Generative photography and AI video

2022 to 2024

- Produced AI video for an AMC documentary.
- Trained custom Stable Diffusion models on people and subjects and placed them in scenes with inpainting.

Smart contracts and Camera Person

2021 to 2022

- Learned smart contract development: wrote and deployed contracts, built the minting sites around them and shipped tools like rarity rankings and sales bots.
- Built Camera Person, a collective art gallery of 122 photographers from 39 countries, with every sale split equally on-chain.

East Coast music venues · House photographer

2018 to 2021

- Shot for venues including Franklin Music Hall in Philadelphia and I.M.P.'s venues in Washington, D.C., then for the artists themselves.

OPEN SOURCE CONTRIBUTIONS

Contributions by Josh Berns / MASHBOX Labs. Full list with links: boshjerns.com/oss

 llama.cpp	 #30737 Exact f32 gelu_quick on CPU, with SIMD.
	 #30736 Qwen-VL image resize that matches transformers.
	 #30108 Exact GELU for ModernBERT encoders.
	 #29692 Configured activation for reranker heads.
	 #29627 Configured pooling for ModernBERT rerankers.
	2 open pull requests
  JAX / XLA	 xla#51365 Wrong TPU results from a dilated convolution. beckerhe (XLA team): "Indeed, this looks wrong. Thanks for reporting this. I'm working on a fix!" Fix: openxla/xla#51380 .
	2 open pull requests, 1 bug report
 MaxDiffusion	4 open pull requests
 Hugging Face	1 open pull request
 Kimodo	1 open pull request
(NVIDIA)	
 SocialNavSUB	1 open pull request
 PyMeshLab	1 open pull request
 Apple Metal	2 bug reports to Apple
 Blender	1 open pull request

SELECTED PROJECTS

- **Mithril Whisper**: private on-device dictation for Mac and Windows. [GitHub](#)
- **Return of the Colossi**: a Shadow of the Colossus homage in the browser, built with AI, three.js, WebGPU, Jolt physics and headless Blender; first playable in 48 hours. returnofthecolossi.com

TECHNOLOGIES

Numbers point to Where I used it. TPU work at MASHBOX Labs is research and engineering, not a public service.

Languages  Rust¹  TypeScript²  Python³  Swift⁴  C and C++⁵  JavaScript⁶  Solidity⁷
Backend frameworks and runtimes  Tokio⁸  Axum⁹  NestJS¹⁰  Prisma¹¹  Node.js¹²
 Deno (Supabase Edge Functions)¹³  Bun¹⁴
Databases and storage  PostgreSQL¹⁵  PgBouncer¹⁶  Redis¹⁷  SQLite¹⁸  PostgREST¹⁹
Messaging and streaming  WebSockets²⁰  Apache Kafka²¹
Infrastructure and cloud  Hetzner²²  Cloudflare²³  Cloudflare Workers and R2²⁴  Cloudflare Pages²⁵
 Google Cloud TPU²⁶  Google Cloud Storage²⁷  Amazon Web Services (ECS, MSK)²⁸  Docker²⁹  Caddy³⁰
 Linux and systemd³¹  Supabase³²  Pulumi³³  Kubernetes³⁴  OpenTofu³⁵
CI/CD and DevOps  GitHub Actions³⁶  Harness³⁷  Playwright³⁸
Observability  Grafana³⁹  Grafana Loki⁴⁰  PostHog⁴¹
Web frontend  React⁴²  Vite⁴³  Tailwind CSS⁴⁴  Next.js⁴⁵  three.js⁴⁶  WebGPU⁴⁷  WebAssembly⁴⁸
 WebRTC⁴⁹
Mobile and desktop  SwiftUI⁵⁰  App Store and TestFlight⁵¹  Apple Foundation Models⁵²
 React Native and Expo⁵³  Tauri⁵⁴  NSIS⁵⁵  WalletConnect (Reown)⁵⁶

ML and AI llama.cpp⁵⁷ JAX⁵⁸ Pallas⁵⁹ XLA⁶⁰ Flax and Optax⁶¹ MaxDiffusion⁶²  PyTorch⁶³
 Hugging Face Hub⁶⁴  ONNX Runtime⁶⁵  NVIDIA Parakeet⁶⁶ Whisper (ExecuTorch)⁶⁷ Stable Diffusion⁶⁸
 Claude Code and Codex⁶⁹ **LLM APIs** (OpenAI, Anthropic)⁷⁰  Model Context Protocol⁷¹
Game, graphics and 3D  Bevy⁷²  Unreal Engine 5⁷³  Blender⁷⁴ Jolt Physics⁷⁵
Blockchain and onchain data  Polygon⁷⁶ Alloy⁷⁷  viem and ethers⁷⁸  Chainlink Data Streams⁷⁹
The Graph (graph-node)⁸⁰ Envio HyperIndex⁸¹ Goldsky⁸² Hyperliquid⁸³
Payments, auth and SaaS  Stripe⁸⁴  Mintlify⁸⁵  Telegram Bot API⁸⁶  Railway⁸⁷

EDUCATION

Self-taught.

PRESS

- [agg.market acquires Polynode, the fastest Polymarket data API](#) (Chainwire, July 2026)
- [agg.market acquires Polynode, a Polymarket data API provider](#) (Prediction News)

WHERE I USED IT

1. Polynode's real-time API and orderbook service.
2. agg.market venue integrations (Betdex, Tip Run, Pred).
3. Polynode's Python SDK for streaming Polymarket data.
4. The native agg.market app for iPhone, iPad and Mac.
5. Reranker and encoder model support contributed to llama.cpp.
6. Cloudflare Workers, such as a data proxy for Polynode.
7. Camera Person contracts that split every sale on-chain.
8. Async runtime for every Polynode Rust service.
9. Polynode's REST and WebSocket endpoints.
10. Venue modules and execution handlers in agg.market's API.
11. agg.market migrations, including a non-blocking index rebuild.
12. agg.market's API and executor, and Polynode's chain indexers.
13. Polynode billing and key-management Edge Functions.
14. One-off data migration jobs at Polynode.
15. Multi-terabyte Polynode databases with 1.2B+ fills.
16. Connection pools in front of Polynode's Postgres.
17. Polynode API keys, rate limits, live market state, streams.
18. Mithril chats and on-device full-text and vector search.
19. Polynode usage metering data for billing and dashboards.
20. Polynode streams of decoded Polymarket orders and trades.
21. agg.market venue integrations on its Kafka pipeline.
22. Polynode's bare-metal servers for real-time ingestion.
23. DNS, TLS and abuse protection for Polynode's public domains.
24. MASHBOX app downloads, gated by the player's pass.
25. mashbox.gg and the Return of the Colossi site.
26. MASHBOX Labs model research on v5p and v6e spot capacity.
27. Research weights, compile caches and queues near the TPUs.
28. agg.market backend services on ECS with Amazon MSK.
29. Polynode's Postgres databases and supporting services.
30. Polynode's TLS edge and blue/green traffic switches.
31. Every Polynode service as a systemd unit, with runbooks.
32. Polynode accounts and billing; self-hosted MASHBOX accounts.
33. agg.market infrastructure code for new venues in production.
34. Local k3d clusters to run and test agg.market's services.
35. Infrastructure-as-code adoption for agg.market's AWS stack.
36. Polynode CI: lint, contract tests, SDK signature tests.
37. agg.market release pipelines with OIDC access checks.
38. End-to-end browser checks of the Return of the Colossi game.
39. Polynode dashboards for latency, per-key usage and capacity.
40. Polynode logs, to tie latency spikes to a customer key.
41. Sign-up and checkout analytics on the Polynode dashboard.
42. The Polynode dashboard for sign-up, API keys and billing.
43. Builds for the Polynode dashboard.
44. Styling for the Polynode dashboard.
45. UI changes to agg.market's shared components for its web apps.
46. Return of the Colossi, a browser game.
47. Rendering for Return of the Colossi.
48. The MASHBOX Rust engine compiled for the browser.
49. Direct player connections for MASHBOX browser multiplayer.
50. The agg.market app UI, adaptive for iPhone, iPad and Mac.
51. Mithril on the App Store; agg.market beta on TestFlight.
52. Apple's on-device model in Mithril, next to open models.
53. The first versions of Mithril.
54. The MASHBOX Windows app that installs and launches the engine.
55. The signed MASHBOX Windows installer.
56. Ethereum wallet sign-in in the agg.market app.
57. On-device GGUF models in Mithril; model support sent upstream.
58. MASHBOX Labs TPU inference and training research code.
59. Custom TPU kernels in MASHBOX Labs research; fixes sent to JAX.
60. Compiling and profiling on TPU in research; compiler bugs reported.
61. Fine-tuning research runs on TPU.
62. Diffusion models on TPU in research; fixes sent upstream.
63. Reference models to check and export weights for TPU ports.
64. Mithril model downloads with memory-fit and SHA-256 checks.
65. On-device text embeddings in early Mithril versions.
66. On-device speech recognition in Mithril.
67. On-device speech-to-text in Mithril and Mithril Whisper.
68. Custom models for AI video in an AMC documentary.
69. Parallel coding agents behind review and verification gates.
70. Quantish agents that traded prediction markets.
71. MCP servers that let Quantish agents read markets and trade.
72. MASHBOX's Rust game engine, mixing worlds with multiplayer.
73. World and simulation prototypes at MASHBOX Labs.
74. Headless asset generation for Return of the Colossi.
75. Climbing and collisions in Return of the Colossi.
76. Real-time decoding of Polymarket activity for Polynode.
77. Event and signed-order decoding in Rust, with parity tests.
78. Order signing in Polynode's TypeScript SDK.
79. Live reference prices for crypto-round markets (Polynode).
80. Self-hosted subgraphs to rebuild Polymarket history.
81. Indexers for live P&L and combo markets (Polynode).
82. Subgraph data for backfills and position cross-checks.
83. An indexer and API for HIP-4 prediction markets.
84. Polynode subscriptions, with webhooks that provision API keys.
85. Polynode's public API docs and changelog.
86. Polynode production alerts.
87. Hosting for Quantish's agent services and databases.