MASON JOEY

Tel 13500038851 mason_joe joe-mason
joey.gq joejoey.ma@gmail.com

Skills

- **Web Development:** React, React Native (Expo), Capacitor, Electron, TypeScript, Tailwind CSS / CSS-in-JS, Redux, Zustand, Zod, Tanstack Query.
- **Backend:** Node.js, Next.js, NestJS (Encapsulated CRUD RESTful APIs), GraphQL, WebSocket, Middleware, Serverless (Vercel, Cloudflare, AWS Lambda/ECS Cold Start Optimization), Edge Distributed Microservices, API Gateway, Redis, Kafka, gRPC.
- **Web Frameworks & UI:** MedusaJS, Better Auth, GSAP Motion Animations, Aceternity / shadcn/ui, AntD.
- **Build Tools & Performance:** Webpack, Vite, Turborepo, Monorepo, pnpm workspaces. performance optimization (code splitting, bundling optimization, asset optimization) tailored to business scenarios.
- **DevOps & Cloud:** Prisma, Drizzle ORM, PostgreSQL/D1, R2/S3. Docker Containerization, Kubernetes (K8s) Clusters, CDN Caching, Load Balancing. Sentry.io (Error Tracking), ELK Stack (Elasticsearch/Kibana). CI/CD / MLOps, Automated Testing (Jest), Canary Releases, A/B Testing. Familiar with Azure, AWS, Google Cloud, Vultr, and Alibaba Cloud.

Work Experience

HKBN (Hong Kong Broadband Network) | *Aug 2023 – Present*

Senior Software Engineer | AI ITBS

- **AI Infrastructure:** Spearheaded the "Zero-to-One" infrastructure deployment of an **AI Portal**, empowering the BS-CS (Business Solutions & Corporate Services) ecosystem.
- **Toolchain Development:** Built a comprehensive AI-driven toolchain integrating APIs, Development workflows, and Vector Knowledge Bases.

Guangdong Lijin Logistics Co., Ltd. | *Feb 2021 – Jul 2023*

IT Supervisor / Lead | Engineering Full Stack Architecture Team

- **ERP Architecture:** Led the Full Stack Architecture Team in refactoring and developing a multi-platform **ERP System** (Web, App, Desktop).
- **System Stability:** Managed high-concurrency data streams and cross-departmental integration. Successfully implemented containerized deployments with robust disaster recovery protocols.

Guangzhou Xinzhongying Cross-border Co., Ltd. | *Feb 2020 – Feb 2021*

Frontend Developer / DevOps

- **E-Commerce Platform:** Developed a cross-border e-commerce platform, implementing comprehensive user event tracking (telemetry), association analysis, and risk control data operations.
- **Optimization:** Focused on system optimization and successfully managed production launches.

Projects

AI Portal – Enterprise AI Data Orchestration & Prompt Platform

Next.js RAG Pipeline Data Lake Kubeflow PyTorch Airflow GSAP

- Led the **0→1 build of an enterprise-grade AI platform**, owning **LLMOps** and a **Microservices-based architecture** integrating retrieval systems, data pipelines, and cost/SLA governance.
- Unified **multi-model and multi-retrieval chains** behind an **API Gateway**, and built **data-driven decision loops** covering Prompt strategies, retrieval hit policies, token cost tracking, re-ranking, latency, and model parameters.
- Integrated **Langfuse** for user feedback and enabled **fully asynchronous end-to-end observability**, with **Query / Vector / Prompt multi-level caching**, plus **A/B testing and canary releases** to ensure traceability and measurable outcomes.

- Achieved **3× overall throughput**, **P95 latency reduced by 60%+**, and **per-request LLM token cost reduced by 40%**.
- Designed and implemented a **Data Lake + RAG Pipeline**, orchestrated via **Kubeflow + Airflow**, enabling continuous ingestion and updates across multiple business domains.
- Processed **TB-scale heterogeneous data**, including **25,000+ artifacts**: structured tickets (tables), manuals (mixed text-image PPT/PDF), handwritten/scanned cases, long-form contracts, and audio (sales calls, customer service recordings, meeting summaries).
- Leveraged **pgvector + OpenSearch** with separated storage, combined with **HNSW + BM25 + RRF hybrid retrieval**, elevating the retrieval system from "usable" to **"verifiable"**.
- **Top-K hit rate improved by 50%**, **context noise reduced by 30%+**, and ingestion cost optimized.
- Supported **40+ core business scenarios** across operations, risk control, analytics, and enterprise knowledge search using a **1M-Menu Agent–based multi-agent architecture with configurable rules**.
- Closed the loop from AI answers → feedback → Q&A iteration, achieving **>95% recall@k with evidence linked responses and consistency guarantees**, reducing average manual query time **from ~10 minutes to <30 seconds**, and lowering repetitive communication costs by **60%+**, significantly improving cross-team efficiency.
- Built a **self-owned Video Surveillance Vision AI pipeline** (Ping An Clock project), covering data labeling, keyframe extraction, **LoRA fine-tuning**, and **model distillation**, with production-grade training workflows and **edge deployment**.
- Reduced **false positive rate by ~45%** (based on 30 days of production monitoring), and achieved **92% accuracy** on a labeled evaluation set with stable recall.
- Delivered the **AI Portal frontend** with high-fidelity **Figma-to-production** implementation and UX animation ownership.
- Built with **Next.js App Router**, supporting **multi-session concurrency**, **streaming responses**, and **multimodal outputs** (chat / image generation / video generation).
- Applied **PPR / SSR, component-level caching, and compiler optimizations** to maintain performance while continuously shipping new AI features.
- Ensured **stable sub-second FCP**, **50%+ reduction in interaction latency**, and systematically resolved **SSR hydration issues**, while delivering complex GSAP-powered interactions.

ERP / Financial–Business Integration / Inventory & Logistics Management SaaS

WebView PWA Service Worker Monorepo Redux Ant Design Apollo Cluster

- Built an enterprise-grade logistics management system using **React + Nest.js + Capacitor + Electron**, deployed across **three client types** (regional/provincial hubs, handheld devices, driver intelligent control terminals).
- Adopted **Monorepo + Microservices** to support **responsive, multi-platform delivery**, addressing performance bottlenecks in **tens-of-millions-ton-scale** transportation scheduling and order dispatch.
- Refactored Redux state management by introducing **request-level caching, retries, and concurrency control**, enabling **real-time synchronization + optimistic updates + offline support under weak networks**.
- Improved business continuity to **~80%** during network instability.
- Leveraged **IndexedDB** with encapsulated pagination, sorting, virtual scrolling, and form validation, plus **event-rate optimization**, reducing **initial render time from 17s to 3s**.
- Implemented real-time vehicle and freight tracking with **map-based navigation and second-level manual adjustments**.
- Delivered Excel/CSV parsing, proof-of-delivery handling, watermarking, and electronic waybills tightly coupled with business logic.
- Enabled **route cost control and approval workflows**, reducing reconciliation cycles from **days to hours**.
- Implemented **four-level address settlement**, last-mile dispatch, order pooling, ride-hailing–style **intelligent consolidation**, and cost-performance metrics visualization—driving **2× overall transportation efficiency** by reducing empty runs and low load rates.
- Designed a **BFF microservice layer** for API aggregation and analytics, implementing **hierarchical rate limiting and circuit breaking via Redis Lua**.
- Used **DataLoader** to resolve **GraphQL N+1 query issues**, significantly reducing DB query counts and response latency.

- Implemented **JWT-based dynamic authorization with a five-layer permission model**, supporting **1000+ concurrent connections** and **600+ vehicles with real-time location streaming**, maintaining **<500ms end-to-end latency**.
- Built **Kafka-based data pipelines** and deployed via **Kubernetes**, with **CI/CD automation**, **Jest + E2E testing**, **Sentry performance monitoring**, and **robust WebSocket heartbeat/reconnection mechanisms** to prevent polling stalls.
- Supported **40+ Kafka topics** with a **high-availability, horizontally scalable cluster architecture**.

Node.js–Based Independent E-commerce Platform with POS & O2O Integration

Express.js ORM react-i18next Stripe ChatWoot

- Developed a full-stack **online–offline integrated (O2O) POS commerce system** using Node.js.
- Implemented **RESTful APIs with MVC architecture and ORM-based CRUD abstraction**, integrating **react-scanner** for barcode-based inventory and checkout workflows.
- Customized **Stripe payment gateway** with webhook-driven callbacks in Express.js.
- At the driver layer, enabled **hardware-level integration** with barcode scanners and receipt printers, completing the **O2O transaction loop**.
- Built a **>90% coverage user behavior tracking system**, leveraging **sendBeacon** for reliable data transmission on page unload.
- Analyzed conversion funnels and behavioral metrics to guide targeted optimization of **LCP, CLS, PV, UV**, and other performance indicators, directly supporting operational decision-making.
- Implemented **i18next-based internationalization** with on-demand language bundle loading, and customized an **intelligent ticketing system based on ChatWoot**, enabling **24×7 automated customer support**.
- Increased overall conversion rate by **28%**.

Education

South China University of Technology

Degree: Bachelor of Computer Science and Technology

Language : Cantonese,Mandarin,English