Le Xuan Cuong

linkedin.com/in/lxcuong | github.com/lexuancuong | cuongkane.com

Email : cuongle.software@gmail.com

Mobile : +84 77 888 7894

Summary

Senior Backend Engineer with 6 years building highly-available, scalable cloud systems processing millions of daily transactions. Grew from individual contributor to technical owner of 4 microservices. Led architecture design for 10+ Technical Designs across microservices, ML systems, and real-time pipelines. Highly adaptive to emerging technologies (AI agents, LLMs, modern tooling). AWS Certified Solutions Architect.

Education

- **HCM University of Science** — Ho Chi Minh City, Vietnam
  Master of Science in Computer Science — *Sept. 2021 – Sept. 2025*
- **HCM University of Science** — Ho Chi Minh City, Vietnam
  Honor Program - Bachelor of Science in Information Technology; GPA: 8.4/10 — *Sept. 2017 – Sept. 2021*
- **Coder School - Data Science Course** — Ho Chi Minh City, Vietnam
  Machine Learning course sponsored by Facebook (Certificated) — *Aug. 2019 – Dec. 2019*

Experience

- **Parcel Perform** — Ho Chi Minh City, Vietnam
  Senior Backend Engineer — *Aug 2024 - Present*
  - **System Architecture & Design**: Led architecture design reviews for 10+ Technical Designs spanning ML systems, analytics platforms, and real-time pipelines. Authored cross-team TD for ML-based carrier identification across 7 microservices, reviewed by CTO.
  - **Microservices & Database Splitting**: Executed 2 microservice decompositions and 2 database splits to improve system modularity and team autonomy. Designed backward-compatible APIs and coordinated cross-squad releases for zero-downtime migrations.
  - **High-Availability Cloud Systems**: Built mission-critical Reports & Analytics platform (Jaspersoft/ClickHouse) and Alert Management system using DDD architecture with Dagster orchestration. Optimized chart loading performance through static file caching, improving customer dashboard experience.
  - **Real-Time Data Pipelines**: Architected ML-based carrier identification pipeline spanning 7 microservices (Flink CoreJobs, Scraper, DataIO, Scheduler), processing millions of daily events. Designed seamless integration with legacy regex flow using feature toggles and Kafka protobuf schemas, improving parcel found rate while preventing recursive scrape loops.
  - **AI/ML Solutions**: Architected L3 AI agents using LangChain/LangGraph with ReAct pattern and AWS Bedrock. Built tool-based architecture for Root Cause Analysis and automated insights generation. Reduced LLM costs by 20% through strategic model selection while maintaining output quality.
  - **Engineering Excellence**: Pioneered AI-assisted Technical Design process (first squad at company). Delivered 105+ tickets as sole backend engineer with 100% on-time delivery. Shipped AI features under tight deadline for CEO conference demo, keeping pace with market competitors.
  - **Release & Deployment Strategy**: Designed release and deployment plans using canary deployments, A/B testing, and feature toggles. Introduced process improvements: 4-day advance deployment booking, mandatory QA sign-off on deployment plans, achieving zero-downtime releases with minimal incidents.
  - **Observability & Sustainable Growth**: Built Grafana dashboards tracking P95/P99 latencies, Kafka consumer lag, and success/failure rates. Championed dashboard-first approach before major changes and debugging. Enabled data-driven capacity planning and sustainable system evolution.
  - **Mentorship & Knowledge Sharing**: Mentored 2 engineers to promotion-readiness through architecture discussions, code reviews, and pair programming. Delivered 15+ Open Space sessions on system design, AI/LangChain, and Flink; fostered bidirectional learning culture to continuously grow team expertise.
  - **Engineering Leadership**: Conducted 6+ technical interviews for junior/mid-level engineers. Established runbook and postmortem culture for incident response. Championed coding standards (PEP8, TDD, Django structure). Led backlog grooming with story point re-estimation techniques for accurate sprint planning.
- **Parcel Perform** — Ho Chi Minh City, Vietnam
  Backend Engineer — *Jul 2022 - Aug 2024*

- **Scalable API Development**: Architected Public Booking API with OpenAPI spec serving multi-carrier integrations (DHL, MyParcel, Shippo). Implemented S3 pre-signed URLs, Kafka async processing, and Celery-based rate limiting to protect third-party APIs.
- **Extensible Integration Architecture**: Designed base booking integration abstraction supporting 50+ carrier integrations. Gathered requirements through practical usage testing and Product collaboration to build flexible, extensible foundation.
- **High-Volume Data Pipelines**: Built and maintained data pipeline processing 5M+ daily events across scheduler, scraper, data I/O, and Apache Flink services. Ensured sub-second latency for mission-critical shipment tracking.
- **Cross-Functional Collaboration**: Authored 4+ Technical Designs collaborating with Product and Design through HLD meetings. Applied ubiquitous language (DDD) to align domains. Negotiated deliverables with Product, presenting trade-offs on resources, timing, and technical constraints.
- **Email Extraction Platform**: Designed pattern-based email extraction engine (Regex/XPath/custom formatters) with Kafka CDC pipeline. System running 3+ years in production processing shipment notification emails.
- **Architectural Decision Records**: Documented architectural decisions with rejected alternatives and trade-off analysis in Technical Designs. Evaluated 4 approaches for email processing; selected app-based CDC, recording rationale for future reference and team alignment.
- **Sync-to-Async Reliability Pattern**: Designed fault-tolerant Booking API spanning 5+ microservices with sync-to-async fallback pattern. Implemented Kafka-based async processing as fallback when synchronous calls timeout, ensuring booking completion despite component failures.
- **Celery Pipeline Architecture**: Built Collection Point data pipeline using Celery canvas patterns (chain, chord, group) with optimized worker configuration for scalability. Designed simple, maintainable task breakdown (8K+ postal codes chunked into parallel sub-tasks). Evaluated 3 scheduling approaches; selected Celery Beat for auto-retry and fault tolerance.
- **System Reliability & Incident Response**: Resolved critical production incidents: BFCM traffic spikes (10x load), diagnosed Redis cluster hotspots. Built recovery script in 24 hours to restore 1.5M records across multiple services for enterprise customer. Documented solutions via technical blog posts.
- **Large-Scale Data Migrations**: Executed production data migrations handling SQL timeout constraints, autovacuum optimization, and strategic indexing. Implemented incremental update patterns to minimize downtime and ensure data consistency across distributed systems.
- **Testing & Developer Experience**: Authored testing best practices guide (AAA pattern, black-box testing, flaky test prevention) presented company-wide. Led tool adoption: Poetry, Ruff/Pyright, Django Ninja. Created onboarding documentation enabling 5 engineers to ramp up faster on complex domain.

- **Cinnamon AI Labs** — Ho Chi Minh City, Vietnam
  Software Engineer — *Aug 2020 - Jul 2022*
  - **Enterprise Client Delivery**: Delivered production systems for Fortune 500 clients (Deloitte, Bridgestone, Daikin, Sumitomo) across document processing and AI automation domains.
  - **Multi-Framework Expertise**: Adapted rapidly across web frameworks (Flask, FastAPI, Django) based on client requirements. Designed RESTful APIs using Repository-Services pattern and SOLID principles.
  - **Multi-Cloud Infrastructure**: Deployed and maintained services across AWS, Azure, and GCP based on enterprise client preferences. Containerized applications with Docker for consistent cross-cloud deployments.
  - **Cross-Team Collaboration**: Partnered with AI researchers and product managers to translate ML models into production-ready solutions meeting client SLAs.

Projects

- **Codlis - AI Code Visualization (Master's Thesis, H2 2025) (codlis.com)**: Multi-agent LLM system for interactive code comprehension. Built 5-agent architecture (Scanner, Analyzer, Architect, Illustrator, Proofreader) with Knowledge Graph and ReAct reasoning. User study: 2.9x task success rate, 29% cognitive load reduction vs baseline (p¡.001). Stack: Django/Celery, Next.js, LangChain.
- **SweatCharge - Sports Club Management (H1 2025) (sweatcharge.com)**: Full-stack member management with role-based SSO (owner/manager/player permissions), fuzzy name matching for player association from raw text, automated cost splitting and billing. React/TypeScript with production users.

Certifications

- **AWS Certified Solutions Architect - Associate** — Amazon Web Services
  Valid: Jan 2025 – Jan 2028

Awards & Recognition

- **Parcel Perform (2022-Present)**: Member of the Month (3x), Best Presentation Award, AVA AI Award for AI adoption, Professional Style Award.
- **Cinnamon AI Labs (2020-2022)**: Employee Recognition Award for outstanding technical contributions.

Technical Skills

- **Architecture**: High-Availability Systems, Microservices, DDD, System Design, RESTful APIs, TDD, Idempotency, Outbox Pattern, Feature Toggles, SSO/OAuth
- **Big Data & Pipeline**: Apache Flink, Kafka, Celery (canvas: chain/chord/group), Dagster, RabbitMQ
- **Infrastructure & Observability**: AWS (EC2, S3, Bedrock), Docker, Grafana, Sentry, Alerting Systems, CI/CD
- **AI/ML**: LangChain, LangGraph, LangSmith, AI Agents, Prompt Engineering
- **Languages**: Python, Scala, Java, TypeScript, SQL
- **Databases & Data Stores**: PostgreSQL, Redis, ClickHouse
- **Web Frameworks**: Django, Django Ninja, FastAPI, Pydantic, Next.js/React (side projects)

Languages

- **English**: B2 Upper Intermediate – Aptis ESOL (British Council) 176/200 **Vietnamese**: Native