



Diego Orue

Date of birth: 09/05/2005 | **Nationality:** Spanish | **Phone:** (+34) 637234888 (Mobile) |
Email address: orueret@gmail.com | **Address:** 48100, Mungia, Spain (Home)

• ABOUT ME

Software engineer with production experience across two demanding systems: real-time video inference pipelines on NVIDIA GPUs, and a document processing and RAG platform handling over five million records. In both cases I have owned the full cycle development in C++ and Python, containerised deployment on Linux, infrastructure as code, and observability with Prometheus and Grafana. My working principle is that critical systems rarely fail with a visible error; they degrade quietly. So I instrument first, measure, change one thing, and measure again. Cloud Native specialisation covering AWS, Kubernetes and infrastructure as code.

• WORK EXPERIENCE

24/02/2025 - CURRENT - VALLE DE TRAPAGARÁN, SPAIN

SOFTWARE ENGINEER - BACKEND, COMPUTER VISION & DATA RADIMIR SEGURIDAD SL

Video inference infrastructure

- Built and integrated real-time video inference pipelines using NVIDIA DeepStream and Triton Inference Server on NVIDIA H200 and Tesla GPU servers, for person and object detection in access control systems.
- Deployed and tuned detection models (YOLOv26, YOLOv11, PeopleNet) for real-world conditions, balancing accuracy against latency and cost per stream.
- Migrated the ingest layer from a CPU-bound MediaMTX setup to the GPU-resident pipeline provided by Savant's embedded media server, keeping frames in GPU memory end to end and cutting per-stream latency significantly.
- Instrumented pipelines with Prometheus and Grafana per-stream FPS, VRAM usage, stream drops and reconnections catching degradations before they reached the service.

Data platform and RAG system (lead developer)

- Designed and built an ingestion, OCR and extraction pipeline over 5M+ public procurement records, backed by PostgreSQL and deployed in containers.
- Instrumenting per-document extraction status cut phantom errors from ~850 to ~22 and surfaced failures that had been happening silently for weeks.
- Built the retrieval and generation layer on self-hosted models (embeddings, reranking, two OCR engines) across 144 GB of GPU memory, plus an analytics layer estimating award probability against bid discount.
- Operated the underlying Linux and Docker infrastructure; version control and code review through pull requests on GitHub.

• SKILLS

Languages:

C++ | Python | Go | Rust | PHP | TypeScript | JavaScript | Java | C#

Backend:

Laravel Framework | Express 5 | Node.js | Inertia.js | Symfony | Prisma | API REST

GPU & machine learning

NVIDIA DeepStream | Savant | Triton Inference Server | Ultralytics | CUDA | PeopleNet | ANPR | TensorRT | Roboflow | RAG pipelines | YOLOv26/YOLOv11

DevOps and Cloud:

AWS - Lambda, ECS, EKS, S3, API Gateway, Serverless | Docker | Kubernetes | Terraform | GitLab CI/CD | GitHub Actions | Nginx | Prometheus + Grafana

DB & Cache:

PostgreSQL | LanceDB | Pgvector | MySQL/MariaDB | MongoDB | Oracle SQL | Redis

Real Time & Messaging:

WebRTC | Laravel Reverb | Socket.io | BullMQ | ZeroMQ | Mediamtx

Integraciones:

Hikvision ISAPI | Milestone XProtect (ONVIF) | Stripe | Firebase | MinIO | Datadis/ESIOS/AEMET

Systems & Ways of working:

Linux (Ubuntu, Red Hat) | Git | Agile development | code review via pull requests | technical documentation

● **PROJECTS**

ConteoBK: Real-time video analytics for occupancy control — Bilbao Kirolak (Bilbao City Council) System development, focused on the video pipeline, the backend and the GPU/Kubernetes infrastructure.

- Multi-camera occupancy counting via line crossing, real-time occupancy dashboard and live video over WebRTC.
- Vision pipeline on multi-GPU Kubernetes (up to 7× NVIDIA Tesla T4, one Savant/DeepStream pipeline per GPU), integration with Milestone XProtect (ONVIF Bridge) and event signing with HMAC-SHA256.
- Stack: Laravel 12, React 19, Savant 0.6 / DeepStream 7.1, YOLOv11 / RF-DETR, Kubernetes, MySQL, Redis, ZeroMQ.

Plusia: Collective self-consumption energy community management platform Lead developer — originally my final-year project (DAM), later refined at the company together with the team on its electrical engineering side.

- Peer-to-peer solar energy trading between prosumers and consumers, with automated monthly settlement of the energy balance.
- Modelling of the Spanish self-consumption regulatory framework (Ley 5/2024, RD-ley 7/2026, RD 244/2019) and integration with external APIs: Datadis, ESIOS/REE and AEMET.
- Stack: Laravel 13 (PHP 8.4), Inertia + React (TypeScript), PostgreSQL, Redis, Laravel Reverb (WebSockets), Sanctum.

28/08/2025 - CURRENT

- End-to-end platform for ingestion, OCR, indexing and natural-language querying of public tenders, running on self-hosted LLM infrastructure (144 GB of VRAM) deployed in containers.

ANPR car park system (Hikvision): Vehicle access control through automatic licence plate recognition

DevOps role : containerising the project and implementing video streaming.

- Full containerisation (multi-stage builds, orchestration of 7 containers) and video streaming implementation with MediaMTX (RTSP/WebRTC/HLS).
 - Integration with Hikvision cameras (ISAPI) and automatic synchronisation of licence plate allowlists per car park.
 - Stack: Laravel 12, React 19, MySQL, Docker, MediaMTX, Supervisor (queue workers).
-
- Multi-tenant SaaS platform with a web dashboard, mobile app and tablet app: clock-in via geolocation, QR and PIN, with cryptographic record integrity through hash chaining.
 - Stack: Express 5 + TypeScript + Prisma, MySQL, Redis, BullMQ, Socket.IO, React 19, React Native, Firebase, MinIO, Stripe.

inventarioIT: IT infrastructure inventory management (clients sites devices) Full-stack development in a team, with solid engineering practices.

- IPAM modules, network topology, Excel import and PDF/Excel export.
- Workflow with Git flow, code review, Conventional Commits, tests (Pest) and CI/CD on a self-hosted runner.

• Stack: Laravel 13, React 19 + Inertia v2, PostgreSQL, Redis, Docker, GitHub Actions.

• LANGUAGE SKILLS

Mother tongue(s): **SPANISH**

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken production	Spoken interaction	
ENGLISH	B2	B2	B2	B2	B2
EUSKERA	B2	B2	B2	B2	B2

• EDUCATION & TRAINING

07/02/2023 - 18/06/2026 - BILBAO

TECNICO DE DESARROLLO DE APLICACIONES MULTIPLATAFORMA (DAM) + ESPECIALIZACION CLOUD NATIVE- CENTRO DE ESTUDIOS ALMI

Website: <https://almi.eus/>

21/02/2024 - 05/06/2024 - BILBAO

DESARROLLO DE VIDEOJUEGOS Y REALIDAD VIRTUAL CON UNITY 3D- LANBIDE

— 300 horas

Field(s) of study: C#, Unity | Website:

https://apps.lanbide.euskadi.net/apps/CE_DETALLE_ESPECIALIDAD?LG=C&ML=FORMEN1&MS=Fcaa&IDCE=IFCD28

2021 - 2023 - MUNGIA

BACHILLERATO- IES MUNGIA BHI

Website: <https://mungia.hezkuntza.net/es/>

• DRIVING LICENCE

Driving licence: B