JEAN-BAPTISTE GIDROL

AI Platform Architect | LLM Inference Engineer

Cyprus / Brussels · +33 6 98 19 25 70 · jbgidrol@hotmail.fr · www.jbgidrol.com

BILLIONS OF TOKENS / WEEK · 15 TEAMS · 24 GPUs · UP TO 99% CACHE HIT · 5M+ USERS

PROFILE

Architect and hands-on engineer who builds and runs self-hosted AI platforms, from GPU model serving to the applications that depend on them. At RTBF, built an in-house inference platform used by 15 teams and serving billions of tokens per week, alongside media AI pipelines, editorial search and personalization for millions of users. Every serving decision is benchmarked; focus on performance, reliability and sovereignty.

EXPERTISE

Inference engineering vLLM, Ray Serve, H100 / L40S, tensor parallelism, FP8 weights and KV cache, FlashInfer, compiled CUDA graphs, prefix-cache-aware routing, EAGLE and DFlash speculative decoding.

AI platform Kubernetes, Ray, LiteLLM gateway, GPU capacity planning, Prometheus, Grafana, Langfuse, GuideLLM and vLLM bench.

Applied AI and data Speech recognition, diarization, OCR, embeddings, RAG, recommendations, Python, SQL, Kafka, Spark, Neo4j, AWS and Azure.

EXPERIENCE

RTBF (client of Amaris Consulting) | AI Platform and Data Engineer

Brussels, Belgium | Mar 2024–present

- Built and run an on-premises inference platform for language, vision, image-generation, embedding and speech models: billions of tokens per week for 15 teams (developers, newsroom, innovation, CRM) behind one LiteLLM gateway with per-team API keys.
- Designed the architecture for 24 GPUs (8 H100, 16 L40S) on Kubernetes and Ray, with L40S replicas as fallback for H100 services and shared observability.
- Built a benchmark suite (GuideLLM, vLLM bench) comparing models, parallelism layouts, FP8 KV cache, vLLM and CUDA versions and speculative decoding, so deployment choices come from data.
- Tuned Qwen3.6 FP8 on a single H100 to about 5,000 tokens/s at 75 concurrent users (roughly 80 tokens/s per user), with no errors.
- Raised prefix-cache hits to up to 99% through cache-aware routing and a shared client harness across the company. Agentic requests that took up to 90 s to start after a model reload now start in under 1 s once warm; production mean time-to-first-token is under 1 s.
- Cut model cold starts from hours to minutes by reusing compilation artifacts and fixing memory-constrained launches.
- Moved media AI to the new GPUs: 12× real time versus 4× before, with a more accurate pipeline (diarization, then language detection and transcription per language).
- Scaled personalization for 5M+ users (2M+ daily recommendations) on 100+ Kubernetes microservices, and built editorial search and RAG over 1.5M articles and years of video transcripts.

ADDITIONAL EXPERIENCE

Amaris Consulting | Domain Manager, AI Inference

Brussels, Belgium | Jan 2024–present | Employer; RTBF is my client engagement

- Technical lead on client calls and adviser on technical offers; upskill consultants through the lecture “Understanding and Running AI in the Real World”.
- Lead internal R&D lab work, including live-sports summarization for beIN Asia and a lakehouse accelerator.

Pragaz | Junior Data Engineer

Lyon, France | Jul 2022–Dec 2023

- Built ETL pipelines from accounting and fleet systems into PostgreSQL on AWS RDS, and Dash applications for commercial and delivery analysis.

SELECTED SYSTEMS

Production GPU inference | RTBF

Serving choices are measured, not assumed: precision, concurrency, throughput, per-user latency and KV-cache behavior compared across model families, vLLM releases and CUDA versions. Compilation artifacts are reused to shorten cold starts.

Best single-H100 result Qwen3.6 FP8, TP1, compiled graph, full FlashInfer. At 75 concurrent requests: 4,970 output tok/s, 12.6 ms median inter-token latency, p95 first token 1.0 s. On long articles (8k in / 1.5k out), 50 concurrent: 4,010 tok/s, p95 first token 2.05 s. Zero errors.

Models in production LLMs: Qwen3.6 35B, Qwen3.8 27B (DFlash), DeepSeek V4 Flash with vision, Mistral (EAGLE). Embeddings: Qwen 0.6B and 4B, Snowflake. Image generation: Z-Image. Speech and OCR: WhisperX, Pyannote, PaddleOCR.

Layouts DeepSeek V4 Flash: TP4 on 4 H100. Qwen3.8 27B: two TP2 replicas on H100 and one TP4 on L40S; L40S replicas are the production fallback behind LiteLLM. Qwen3.6 previously ran as four TP1 replicas on H100 and two TP2 on L40S.

Multimodal media control plane | RTBF

PostgreSQL-backed job system for media AI: idempotent requests, capacity-aware dispatch, attempt leases, retries, cancellation fencing, progress and ETA tracking, ordered callbacks, validated outputs and atomic artifact publication. Connected to Ray Serve services for speech, diarization, transcription and OCR.

Technology Python, PostgreSQL, Ray Serve, WhisperX, Pyannote, PaddleOCR, Kubernetes.

EDUCATION

Master’s in Data Science | OpenClassrooms / CentraleSupélec | 2023–2024

Bachelor’s in Data Analysis | OpenClassrooms / ENSAE | 2022–2023

Bachelor’s in Video Editing | University of Rennes 2 | 2014–2015

LANGUAGES

French and English: fluent | Chinese and Japanese: B1