CLEMENT PALEZIS

Talence / Bordeaux, France | 07 60 40 08 11 | palezis.c@gmail.com
LinkedIn: linkedin.com/in/cpalezis | GitLab: gitlab.com/Nhkp | Portfolio: clement-palezis.dev

HPC / GPU / AI Software Engineer

CUDA | Heterogeneous Computing | ML & LLM Workloads | Backend APIs | Observability | Energy-Efficient Computing

SUMMARY

Software engineer combining high-performance computing, GPU acceleration and applied AI. Experienced in CUDA-based scientific workflows, heterogeneous architectures, HPC workload analysis, ML/LLM benchmarking, backend services, APIs and monitoring dashboards. I bring a profile able to connect low-level performance engineering, AI workload understanding and product-oriented platform tooling for cloud, HPC, AI infrastructure, R&D and performance-oriented software teams.

HPC / GPU / AI SKILLS

HPC & Performance Computing: C, C++, CUDA, OpenMP, MPI, SIMD, CMake, Bash, heterogeneous clusters, task/parallel programming, simulation workloads, benchmarking, profiling, performance and energy metrics.
GPU Acceleration: CUDA programming, GPU kernels, tensor-based methods, GPU-accelerated 3D template matching, scientific computing pipelines, compute-heavy workflow optimization.
AI, ML & GenAI: PyTorch, Scikit-learn, ML pipelines, LLMs, RAG, agents, embeddings, semantic search, LangChain, LangGraph, Hugging Face, OpenAI API, Mistral API, Gemini API.
Backend, MLOps & Observability: Python, FastAPI, REST APIs, Docker, Streamlit, Grafana, MLflow, Evidently, Airflow, CI/CD, model deployment, model monitoring, drift analysis, dashboards.

HPC / GPU / AI PROJECTS

RAG-Based Cultural Recommendation Assistant

Product-oriented prototype | LangChain, FAISS, FastAPI, Streamlit, OpenAgenda, LLM APIs

- Built a RAG-based cultural event recommendation assistant powered by OpenAgenda data and natural-language queries.
- Implemented event ingestion, embeddings, FAISS vector retrieval and response generation with LangChain to recommend relevant events based on user preferences and context.
- Exposed the assistant through a Streamlit test interface and FastAPI-ready architecture to evaluate prompts, retrieval quality, latency and user interactions.

HPC Workload Performance & Energy Analysis

Python, APIs, dashboards, ML pipelines, benchmarking frameworks

- Built tooling to analyze HPC workload performance and energy metrics, identify inefficiencies and support optimization of compute operations.
- Developed evaluation workflows and benchmarking frameworks for LLMs and large-scale simulation workloads.

GPU-Accelerated 3D Template Matching

CUDA, tensor-based methods, cryo-electron tomography, AI-assisted segmentation

- Developed GPU-based methods that reduced tomography processing time from days to minutes.
- Combined GPU acceleration, tensor-based approaches and AI methods for state-of-the-art scientific image processing.

WORK EXPERIENCE

R&D Engineer - Denergium

November 2024 - November 2025 | Energy-efficient computing, ML pipelines, backend services

- Built Python tools, backend services, APIs and dashboards to analyze HPC workload performance and energy metrics.
- Developed ML analysis pipelines and benchmarking frameworks for LLMs and large-scale simulation workloads.
- Designed evaluation workflows to measure workload performance, identify inefficiencies and support continuous optimization for more efficient compute operations.
- Collaborated on applied research involving AI, data processing and sustainable computing.

Software Engineer - Scalian DS / Thermo Fisher Scientific

September 2022 - November 2024 | R&D engineering, GPU acceleration, cryo-electron tomography workflows

- Worked as an R&D engineer on cryo-electron tomography workflows for Thermo Fisher Scientific.
- Developed GPU-based 3D template matching approaches that reduced tomography processing time from days to minutes.
- Built state-of-the-art template matching approaches, including GPU-accelerated, tensor-based and AI-based methods.
- Applied the Segment Anything Model to nanometer-scale tomography segmentation with Universidad de Murcia and the Max Planck Institute.

Software Engineer Intern - CEA

March 2022 - August 2022 | Heterogeneous architectures, simulation code

- Developed software around programming models for heterogeneous architectures in a simulation code.
- Worked on numerical simulation, data visualization, parallel/task programming and heterogeneous cluster computing.

EDUCATION

Master's Degree - AI Engineering | OpenClassrooms | 2026 - Present
Engineering and Data Science Expert - ML optimization, model lifecycle management, model performance, CI/CD deployment for ML applications.
Master's Degree - Computer Science | Universite de Bordeaux / ENSEIRB | 2020 - 2022
High Performance Computing and Data Science - architectures, programming models, HPC and ML algorithms, execution and visualization environments.
Bachelor's Degree - Computer Science | Universite de Bordeaux | 2019 - 2020

LANGUAGES & BEHAVIORAL SKILLS

French: native | English: professional fluency | adaptability, curiosity, collaboration, empathy, continuous learning.