

NATHAN CASSEREAU

AI DevTech @ NVIDIA

 nathan@cassereau.fr

 +33 6 77 54 14 95

 French nationality

 28 years old

 Clamart, FRANCE

 nathan-cassereau

 ncassereau

 cassereau.fr



Senior HPC/AI engineer with a track record in **GPU kernel optimization** (CUDA, Triton: **19× speedup**, **5× memory reduction**), **distributed training** at scale, and ML research collaboration on national supercomputers (**Jean Zay, GB200 NVL72**). Joining **NVIDIA** in October 2026 as AI DevTech.

EXPERIENCE

AI DevTech

NVIDIA

 October 2026 (ongoing)

 Courbevoie, FRANCE

Joining NVIDIA's AI Developer Technologies team, focused on GPU/AI ecosystem enablement for research and industry.

Senior HPC AI Engineer

AMIAD

 October 2025 – October 2026

 Palaiseau, FRANCE

- **Kernel Optimization:** Developed custom **CUDA** and **Triton** kernels achieving significant performance improvements:
 - Second-order optimizer: **19× speedup** and **5× memory reduction** vs. PyTorch on B200
 - Custom CUDA kernel: **4.5× faster** than JAX/XLA compiler
- **HPC Infrastructure Support:** Supporting researchers on **NVIDIA NVL72** supercomputer (Spack, Slurm scripting, cluster management)
- Built **OhMyJob**, a real-time monitoring platform for Slurm HPC jobs (CPU/IO/GPU metrics collected on compute nodes, centralised into dashboards for researchers)
- Infrastructure deployment and automation using **Docker**, **Ansible**, and **OpenStack** for research computing environments.

ML Engineer (Freelance)

Entalpic

 June 2025 – September 2025

 Paris, FRANCE

I optimized their use of **national supercomputers**, achieving **5× throughput improvement** in distributed training. I also implemented specialized **MLIP benchmarks** and optimized training pipelines for chemistry AI models.

Artificial Intelligence Engineer

IDRIS (CNRS)

 May 2021 – May 2025

 Orsay, FRANCE

- **AI support team:** supporting **2000+ researchers** on **Jean Zay supercomputer**, benchmarking hardware accelerators (NVIDIA/AMD GPUs, GraphCore IPU), developing tools for **distributed training**
- **Research collaborations** (PNRIA): AION (Flatiron Institute, **foundational model** for astrophysics), AutoSynth (LaBRI, **LLM code generation**)
- **Teaching:** Deep Learning courses (Intro, **Optimization for Supercomputers**, **LLM Specialization**), NVIDIA bootcamps, JSALT workshops

EDUCATION

MSc. Artificial Intelligence & Deep Learning with Honours

Imperial College London

 2019 – 2020

 London, UNITED KINGDOM

Master in Engineering

CentraleSupélec

 2017 – 2020

 Gif-sur-Yvette, FRANCE

LANGUAGES

English ● ● ● ● ●

French ● ● ● ● ●

SKILLS & TECHNOLOGIES

- GPU & Kernel Optimization (CUDA, Triton, GPU Profiling: Nsight / Torch)
- Distributed & Parallel Computing (OpenMP, PyTorch DDP/FSDP, DeepSpeed, Megatron)
- ML Frameworks & Tools (PyTorch, JAX, HuggingFace, vLLM, W&B, MLflow)
- Deep Learning (NLP & LLM, Computer Vision, Diffusion, Multi-modal, Post-training)
- HPC Infrastructure (Slurm, Spack, Docker, Apptainer, Modules, Ansible, Terraform)
- Programming & Tools (Python, C++, Rust, Jupyter, Git, CI/CD, $\LaTeX$, Grafana)