Hugo AFFATICATI

haffaticati@gmail.com
haffaticati.github.io
https://www.linkedin.com/in/hugo-affaticati/

Technical Experience

Microsoft Seattle, WA, USA

Senior Cloud Infrastructure Engineer **Sep 2025 – current**

- Delivered record-setting performance of 865,000 and 1.1M tokens/sec for LLAMA 70B Inference on GB200 and GB300 racks, featured in Satya Nadella's Build and Ignite keynotes and acknowledged by NVIDIA at annual conference GTC.
- Closed a critical 20% performance gap across 14 AI training models (NeMo framework) within three weeks—after six months of stalled progress—by leading 50+ engineers from Microsoft and NVIDIA through full stack root cause analysis.
- Owned end-to-end MLPerf Training submission on 512 H200 GPUs, architecting scalable infrastructure and delivering 28% cost savings; published results in peer-reviewed white paper, managed vendor and partner engagement.
- Produced competitive intelligence reports to Microsoft SLT analyzing 1k accelerator performance data points used for partner strategy, investing confidence, earnings explanations, and seven new $M+ customer contracts.
- Mentored five full-time employees and two interns (accepted return offers) as Technical Lead for the workloads team.

Cloud Infrastructure Engineer II **Mar 2024 – Aug 2025**

- Developed and launched the AI Benchmarking Guide, standardizing cloud AI infrastructure performance (GPUs, network, storage, etc.) across hardware vendors; led a cross-company team of 15 engineers to drive industry adoption.
- Reduced generative AI inference latency by 46% on LLAMA models by optimizing model configuration (memory, batch size, and KV cache, etc.) to fully leverage next-gen GPU architecture and reduce cost for enterprise customers
- Showcased Azure's AI leadership via two invited technical talks (NVIDIA GTC, SC24) and three on-demand videos

Technical Program Manager II **Sep 2023 – Feb 2024**

- Managed the Microsoft–NVIDIA program that set the LLM training scale record on Eagle, driving planning, testing, and release materials across research, product, and engineering; demonstrated performance gains in production conditions.
- Assessed Microsoft's latest H100-based virtual machines with NVIDIA benchmarks, including NeMo Megatron and MLPerf inference & training, to establish performance standards and total cost of ownership for customers in Generative AI
- Optimized performance for key AI clients, accelerating training and reducing costs up by six through latest code and software stack implementation using Python
- Influenced marketing strategies by creating 10+ technical blog posts and tutorials, accumulating over 20k views

Technical Program Manager I **Sep 2022 – Aug 2023**

- Executed the first-ever public proof of concept for scale training LLMs (GPT-3, 530B parameters) in the cloud using NVIDIA NeMo framework - catalyzing the global rush for AI infrastructure including OpenAI's move to Azure
- Reached sub-two-min BERT Model training by implementing flash attention mechanism with Stanford AI Research Group
- Headlined key conferences Microsoft Build (most attended technical talk) and SC22 about AI infrastructure capabilities

Program Manager **Aug 2021 – Aug 2022**

- Led Microsoft's MLPerf Training and Inference submissions by optimizing Linux and Python ML code on A100 GPUs demonstrating unmatched performance, latency, and accuracy on virtual machines
- Benchmarked four GPU generations across ML and DL models, showcasing 2X performance and cost gain per generation
- Decoded AI workloads through writing 15+ technical blog posts and documentation (approx. 35k views) for beginners

KARL Tech (startup) Bordeaux, France

Co-founder, Entrepreneur, CEO **Mar 2019 – Apr 2020**

- Coded color rendering software for online retail within three months of incubation at Station F (Europe leading incubator)
- Earned first prize in Innovation among 48,000 students at Paris-Saclay University for Technical potential
- Exhibited and pitched to decision makers at CES 2020 (world's leading tech tradeshow) with France's official delegation

Leadership Experience

Microsoft, *Chair of the ERGs for LGBTQ+ employees at Microsoft (Southeast and Azure Core)* **Mar 2022 – current**

- Managed a 13-person leadership team and a global six-person board with bi-weekly meetings and dedicated mentorship
- Tripled annual budget to $30,000 by starting strategic partnerships between Microsoft's ERGs (geo and business based)
- Built a community with quarterly morale events and monthly safe-space meetings to improve D&I culture at Microsoft
- Received Microsoft's Leadership Award for cultural impact in 2023 (highest membership growth and budget growth)

Education

Yale University, Graduate School of Arts and Sciences – MS in Applied Physics 2020-2021

Research: Magic State Fidelity for Quantum Computation Optimization with Prof. S. Puri
Improving precision for Optical and Quantum Electronics with entangled photons with Prof. P. Rakich
Adapting Maxwell's equation for a general laser theory with Prof. A. Stone, Deputy Director of Yale Quantum Institute

Paris-Saclay University, Institut d'Optique Graduate School – BS and MS in Engineering & Quantum Physics 2018-2020

BS GPA: 3.97/4.00 and MS GPA 4.00/4.00, Mentored by Nobel Prize-winning physicist Gérard Mourou
Relevant coursework: Atomic Physics, Quantum Mechanics, Fourier and Non-linear Optics, Experimental Research

PSL University, Université Paris-Dauphine – BS in Applied Economics 2019-2020

Bachelor in Applied Economics, condensed into one year, for students of Top Engineering Schools
Relevant coursework: Macroeconomics, Microeconomics, Econometrics, and International Economics