

Bruno Magalhaes High Performance Computing and Machine Learning

✉ brunomaga@gmail.com 🌐 brunomaga 💬 brunomaga 🔗 <https://brunomaga.github.io> 🏠 Lausanne, Switzerland
🔧 Python, PyTorch, Triton, DeepSpeed. C, C++, CUDA, HPX, MPI 🇵🇹 Portuguese-Swiss 🌍 fluent in English, French, Portuguese, Spanish

Work Experience

present	Principal Researcher for AI Systems and Networks, Huawei Research, Zurich, Switzerland <ul style="list-style-type: none">Started, grew and led the AI Data Center Networks team (7), defining research direction and leading technical implementation.Researching distributed kernel fusion (scale-up) and in-network computing (scale-out), towards next-gen AI network switches.Developed (C++, CUDA, nvshmem) lighter and faster ML kernels via fusion of computation and communication for Megatron-LM Tensor Parallelism, MoEs dispatch+combine, expert sharding, and (ongoing) sequence-parallel attention prefill/decode.
May 2025	Senior Research Engineer for Large-scale AI, Synthesia, Zürich, Switzerland <ul style="list-style-type: none">Started, grew and led the ML performance team (4), aimed at large-scale training and efficient inference of diffusion and transformer models for text, audio and video (Wan, CogVideoX, DALL-E, DDPM, AnymateAnyone). Implemented data, context/sequence parallelism (Ulysses, Ring attention), sharding and activation checkpointing. Performance modelling (scaling laws).Model speedup via kernels fusion (triton, CUDA, torch compile), quantisation (int8 Sage attention), and channels-last format.Created the DeepSpeed modules for distributed curriculum learning setup (PR 5129) and variable batch size and LR (PR 7104).
Nov 2023	
Oct 2023	AI Resident » Researcher » Senior Researcher, Microsoft Research, Cambridge, UK <ul style="list-style-type: none">as Sr Researcher, 2022-23: Efficient inference of vision models on optical hardware with pipeline and model parallelism (C++). Large-scale ML training via data parallelism, sharding, gradient accumulation, activation checkpointing, IO offloading and distillation (DeepSpeed/ZeRO and Torch Distributed/RPC). Distributed Mixture of Experts. Likelihood estimators, information encoding and error correction (LDPC) for storage systems. Gaussian Processes for hyperparameter finetuning. Mentoring of junior members and PhD interns. MLOps on cluster and cloud environments (docker, builds, CI).as Researcher, 2021: computer vision models for thousand-object classification on 3D glass at Project Silica. Presenter of talks on the topics of <i>CPU/GPU optimization, distributed algorithms and AI SuperComputing</i>.as AI Resident, 2019-20: RNNs, GRUs, Encoder-Decoders, and Bayesian Optimization for regression on time series, to improve load balancing of Exchange email servers on distributed exabyte-scale COSMOS databases. Graph Neural Nets for a recommendation system on a trillion-edge graph of meetings, documents, emails and users, stored on a distributed spark database.
Sep 2019	
Aug 2019	HPC engineer » PhD candidate » postdoc researcher, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland <ul style="list-style-type: none">as postdoc: supervision of PhD students and engineers. Scientific reviewer for <i>SuperComputing, IPDPS, and ISC</i> conferences.as PhD candidate, 2015-19: research, development (C, C++, HPX) and publication on the field of distributed-parallel asynchronous variable-order variable-step simulation of spiking neural networks, on Cray and SGI supercomputers (10K+ nodes).Technologies: asynchronous runtime system (HPX) with distributed memory with global addressing (InfiniBand, RDMA, PGAS), and distributed control objects (concurrency, scheduling); dynamic load balancing; vectorization; cache optimization.Teaching assistant (400h) for <i>Unsupervised and reinforcement learning, Project in neuroinformatics and In silico neuroscience</i>.as HPC Engineer, 2011-15: research, development (C, C++, MPI, OpenMP) and publication of methods for parallel/distributed volumetric decomposition, load balancing, spatial indexing, sorting, IO, sparse matrix transpose, and graph navigation, that underlie an efficient processing of neural networks on SGI and IBM Blue Gene Q supercomputers with 16K+ compute nodes.
Mar 2011	
Feb 2011	Junior Architect for IT infrastructures, Noble Group, London, New York, & São Paulo <ul style="list-style-type: none">Design and configuration of Linux servers, CISCO networks, and backup/redundancy sites for physical trading of commodities.
Sep 2009	
Oct 2008	Analyst programmer, Investment Property Databank (now MSCi Real estate), London, UK <ul style="list-style-type: none">Development of a search engine app (C#) and backend (C++) for efficient storage and analytics of financial data.
Mar 2007	

Education

Jun 2019	PhD Computational Neuroscience, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Mar 2015	➢ Thesis <i>Asynchronous Simulation of Neuronal Activity</i> nominated for the EPFL PhD excellency award and the IBM Research award for best thesis in computational sciences. Visiting researcher at Indiana University (US) during Summers of 2015, '16 and '17.
Sep 2009	MSc Advanced Computing, Imperial College London, UK
Oct 2008	➢ Grade: Merit. Thesis <i>GPU-enabled steady-state solution of large Markov models</i> awarded distinction and published at <i>NSMC'10</i> .
Jul 2007	BSc Systems Engineering and Computer Science (5 year degree), University of Minho, Portugal
Oct 2002	➢ Grade: A, top 10%. Exchange student at the University of Maribor, Slovenia, 2005/06. Intern at IBM. Part-time project at CERN.

Selected Publications [full list on scholar.google.com/citations?user=pirWLLgAAAAJ](https://scholar.google.com/citations?user=pirWLLgAAAAJ)

- 2025 Laser writing in glass for dense, fast and efficient archival data storage, *Nature* (accepted, to be published soon)
2024 Project Silica: sustainable cloud archival storage in glass, *Frontiers in Ultrafast Optics: Biomedical, Scientific, and Industrial Applications XXIV*
2023 Project Silica: Towards Sustainable Cloud Archival Storage in Glass, *SOSP '23: Proc. of the 29th Symposium on Operating Systems Principles*
2020 Fully-Asynchronous Fully-Implicit Variable-Order Variable-Timestep Simulation of Neural Networks, *Proc. ICCS 2020*, Amsterdam, Holland
2019 Exploiting Implicit Flow Graph of System of ODEs to Accelerate the Simulation of Neural Networks, *Proc. IPDPS 2019*, Rio de Janeiro, Brazil
2015 Reconstruction and Simulation of Neocortical Microcircuitry, *Cell* 163, 456–492.