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 – July 2025 | **Principal Researcher for AI Systems and Networks, Huawei Research, Zurich, Switzerland**
- Started, grew and led the AI Data Center Networks team (7). Defined the research direction and implementation of distributed kernel fusion (scale-up) and in-network computing (scale-out), towards next-gen AI network devices on HPC networks.
- Developed (C++, CUDA, nvshmem) fused asynchronous communication-computation mega-kernels for Ring/Tree Attention, Megatron-LM Tensor Parallelism, LLMs, MoEs dispatch+combine, and expert sharding. Many DeepSeek-like improvements.

May 2025 – Nov 2023 | **Senior Research Engineer for Large-scale AI, Synthesia, Zürich, Switzerland**
- 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 and 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).

Oct 2023 – Sep 2019 | **AI Resident » Researcher » Senior Researcher, Microsoft Research, Cambridge, UK**
- 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 *CPU/GPU optimization*, *distributed algorithms* and *AI SuperComputing*.
- 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.

Aug 2019 – Mar 2011 | **HPC engineer » PhD candidate » postdoc researcher, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
- as postdoc: supervision of PhD students and engineers. Scientific reviewer for *SuperComputing*, *IPDPS*, and *ISC* 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 *Unsupervised and reinforcement learning*, *Project in neuroinformatics* and *In silico neuroscience*.
- 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.

Feb 2011 – Sep 2009 | **Junior Architect for IT infrastructures, Noble Group, London, New York, & São Paulo**
- Design and configuration of Linux servers, CISCO networks, and backup/redundancy sites for physical trading of commodities.

Oct 2008 – Mar 2007 | **Analyst programmer, Investment Property Databank (now MSCi Real estate), London, UK**
- Development of a search engine app (C#) and backend (C++) for efficient storage and analytics of financial data.

Education

Jun 2019 – Mar 2015 | **PhD Computational Neuroscience, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
- Thesis *Asynchronous Simulation of Neuronal Activity* 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 – Oct 2008 | **MSc Advanced Computing, Imperial College London, UK**
- Grade: Merit. Thesis *GPU-enabled steady-state solution of large Markov models* awarded distinction and published at *NSMC'10*.

Jul 2007 – Oct 2002 | **BSc Systems Engineering and Computer Science (5 year degree), University of Minho, Portugal**
- 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**

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
2022 Cloud-Scale Archival Storage Using Ultrafast Laser Nanostructuring, Conf. Lasers and Electro-Optics Technical Digest Series 2022
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.