Bruno Magalhaes

Machine Learning, High Performance Computing and Big Data

brunomaga@gmail.com | +41 77 487 89 92 | C, C++, CUDA, HPX, MPI, Python, PyTorch
https://brunomaga.github.io | brunomaga | brunomaga | Portuguese and Swiss
Lausanne, Switzerland | fluent in English, French, Portuguese, Spanish | waterpolo, skiing

Work Experience

present – Sep 2019 | AI Resident » Researcher » Senior Researcher, Microsoft Research, Cambridge, UK

- as Sr Researcher, 2022-: porting Transformer-based models to confidential optical hardware (ongoing). Large ML models scaling via model parallelism, sharding, pipelining, gradient accumulation, checkpointing, IO offloading, shared memory, mixed precision, model compression, and distillation. Likelihood estimators, Multi-armed bandit and Gaussian Processes for error quantification and fine-tuning of optical systems. Information encoding (Gray, Huffman), error correction (LDPC) and channel capacity (Blahut-Arimoto) for non-binary systems. Mentoring of junior members and PhD interns.
- as Researcher, 2021: computer vision models for thousand-object classification on 3D glass at Project Silica. Distributed data parallelism. 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 databases.
- always: full-stack MLOps and CI/CD for cluster and cloud (AzureML) environments. Fine-tuning ML for hardware (network, memory) and business specifications (cost vs accuracy vs runtime trade-off). Performance modelling and analysis at scale.

Aug 2019 – Mar 2015 | PhD candidate » postdoctoral researcher, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

- Research, development (C, C++) and publication on the field of distributed-parallel asynchronous variable-order variable-step simulation and optimization of detailed spiking neural networks, on Cray and SGI supercomputers with 10K+ compute nodes.
- Technologies: C, C++, distributed asynchronous runtime systems (HPX-5) for communication (InfiniBand, RDMA) and computation (concurrency, scheduling); global memory addressing; dynamic load balancing; vectorization; cache optimization.
- Teaching assistant (400h) for *Unsupervised and reinforcement learning*, *Project in neuroinformatics* and *In silico neuroscience*.
- Scientific reviewer for *SuperComputing*, *IPDPS*, and *ISC* conferences. As postdoc: supervision of PhD students and engineers.

Feb 2015 – Mar 2011 | Research Engineer for High Performance Computing, Blue Brain Project, EPFL, Lausanne, Switzerland

- Research, development (C, C++, MPI, OpenMP) and publication of methods for parallel/distributed volumetric spatial decomposition, load balancing, spatial indexing, sorting, I/O, sparse matrix transpose, and graph navigation, that underlie an efficient storage and processing of neural networks on SGI and IBM BlueGene 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 and web/windows app (C#, 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 best thesis in computational sciences award. Visiting researcher at CREST at Indiana University (US), Summers 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 and CERN.

Selected Publications

full list on scholar.google.com/citations?user=pirWLLgAAAAJ

2023 Multi-dimensional optical data writing techniques for cloud-scale archival storage, Proc. Laser Applications in Microelectronic and Optoelectronic Manufacturing (LAMOM) XXVIII

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. International Conference on Computational Science (ICCS 2020), Amsterdam, Holland

2019 Asynchronous SIMD-Enabled Branch-Parallelism of Morphologically-Detailed Neuron Models, Frontiers in Neuroinformatics

2019 Exploiting Implicit Flow Graph of System of ODEs to Accelerate the Simulation of Neural Networks, Proc. International Parallel & Distributed Processing Symposium (IPDPS 2019), Rio de Janeiro, Brazil

2019 Fully-Asynchronous Cache-Efficient Simulation of Detailed Neural Networks, Proc. International Conference on Computational Science (ICCS 2019), Faro, Portugal

2015 Reconstruction and Simulation of Neocortical Microcircuitry, Cell 163, 456–492.