Bruno Magalhaes — High Performance Computing and Machine Learning

email: brunomaga@gmail.com linkedin: brunomaga github: brunomaga webpage: https://brunomaga.github.io
tools: C, C++, CUDA, HPX, MPI, Python, PyTorch, DeepSpeed phone: +41 77 487 89 92 nationalities: Portuguese and Swiss
address: Lausanne, Switzerland languages: fluent in English, French, Portuguese, Spanish hobbies: waterpolo, skiing

Work Experience

present – Nov 2023 | **Senior Research Engineer for Large-scale AI, Synthesia, Zurich, Switzerland**
- Started and been growing a team for large-scale AI applied to Diffusion and Transformer models for text, audio and video. Multi-GPU model scaling via data (DDP, FSDP), activation, pipeline and sequence parallelism. Model speedup via quantization, kernel optimization. Contributed to DeepSpeed with distributed curriculum data loader, and variable batching and LR scheduler.

Oct 2023 – Sep 2019 | **AI Resident » Researcher » Senior Researcher, Microsoft Research, Cambridge, UK**
- as Sr Researcher, 2022-2023: porting of Large Language Models to optical hardware (C++). Large Mixture of Experts. ML models scaling via data parallelism, sharding, pipelining, gradient accumulation, activation checkpointing, IO offloading, mixed precision, and distillation (DeepSpeed/ZeRO and Torch Distributed/RPC). Likelihood estimators, information encoding, error correction (LDPC) and channel capacity (Blahut-Arimoto) for noisy non-binary storage systems. Application of simple genetic algorithms and Gaussian Processes for finetuning of parameters. 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 pipelines for cluster and cloud environments. Performance modelling and finetuning at scale.

Aug 2019 – Mar 2011 | **HPC engineer » PhD candidate » postdoc researcher, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
- as HPC Engineer, 2011-2015: 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.
- as PhD candidate, 2015-2019: 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*.
- Scientific reviewer for *SuperComputing*, *IPDPS*, and *ISC* conferences. As postdoc: supervision of PhD students and engineers.

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. 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, volume 12875, pages 84-89
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. International Conference on Computational Science (ICCS 2020), Amsterdam, Holland
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.