Francesco **Fusco, PhD**

SENIOR RESEARCH SCIENTIST AND SOFTWARE ENGINEER

(+41) 787902996 | fusco@ntop.org | fuscof.ntop.org | https://www.linkedin.com/in/fuscof/ | Francesco Fusco

Scientist and software engineer with a strong background in distributed systems and ten years of experience in machine learning and natural language processing (NLP). Strong and consistent track record of delivering innovative solutions to complex data challenges with granted patents in efficient ML, indexing, compression and networking. Open-source contributor and passionate about performance, I delivered the telemetry layer for storage products used by many Top100 companies and with deployments as large as 50k nodes. I wrote cyber-security solutions scaling to nation-wide deployments. My experience ranges from low-latency network programming in the Linux kernel, to low-level code optimization using SIMD instructions, up to designing novel machine learning model architectures for NLP and recommender systems.

Research interests

Generally interested in everything performance related, especially when it comes to store, index and extract insights from big data. I moved from embedded systems, to large-scale distributed systems, to HPC clusters and from network packets, to matrices, to model architectures. I always found my ways to make things cheaper, faster, or possible by combining a truely full-stack background. Topics (top down order): efficient ML/NLP, LLM alignment, weak/self-supervision, indexing and compression, high-speed packet processing, network/system monitoring, distributed systems, hardware-software codesign.

Professional experience

IBM Research - AI *www.zurich.ibm.com*

SENIOR RESEARCH SCIENTIST - AI SYSTEMS *April. 2014 - now*

- Tech lead for efficient ML - NLP. My mission is to deliver cost-effective and yet high-throughput/low-latency models to perform NLP tasks on extremely large text corpora, i.e., the ones commonly used for pre-training. Some achievements: 1. an unsupervised term encoder matching pre-trained sentence encoders in quality while being *5x* smaller and *10x* faster, 2. models for sequence tagging tasks that are *1-2 MB*, offer *1-2 ms* of latency on a *single* CPU core while providing 99.9% of the accuracy of mBERT.
- Tech lead for LLM alignment - SQL. My mission is to improve IBM proprietary LLMs part of the *IBM Watsonx.ai* offering in data management tasks, such as Text2SQL. Further information can be found in the *FlowPilot* project page, *demoed* in NIPS 2023. Responsible of delivering the fine-tuning pipelines and instruction tuning datasets (via annotations and synthetic data generation).
- Main architect and top contributor of a distributed system to train, index and query large matrices of embeddings built over Terabytes of text. The system has been build from the ground up in C++ using hand vectorized code for x86 (and no ML frameworks).
- Design and implementation of large-scale distributed systems for business intelligence extracting insights from streams of news. Delivered key components (protected by patents) in information extraction, linking, classification, recommendation and QA.
- Internal evaluator for all the patents within IBM about information extraction.
- Research outcome: 7 conference papers, 2 demo papers, 9 patents (3 granted, 6 pending).

Red Hat *www.redhat.com*

SENIOR SOFTWARE ENGINEER - KERNEL NETWORKING *July 2013 - February 2014*

- Maintained the networking stack of Red Hat Enterprise Linux, with focus on network virtualization.
- Contributed upstream to the Linux Kernel and to Open vSwitch (both user and kernel space).
- Designed efficient algorithms for packet filtering (2 patents granted).
- Research outcome: 2 patents granted.

Swiss Federal Institute of Technology (ETH) *www.tik.ee.ethz.ch*

SENIOR RESEARCHER - COMMUNICATION SYSTEMS GROUP *August 2012 - June 2013*

- Exploiting GPUs as indexing co-processors in network appliances: accelerated packet indexing by 20x on consumer GPUs.
- Designed a compression algorithm tailored for networking that offers partial decompression capabilities.
- Research outcome: 4 conference papers.

IBM Research *www.zurich.ibm.com*

PREDOC RESEARCHER - SYSTEMS MANAGEMENT GROUP *March 2009 - July 2012*

- Developed the performance monitoring infrastructure of the *IBM SONAS* storage product. The optimized timeseries database has been later used to enable the monitoring console of other flagship products such as IBM SVC and Storwise V7000.
- Developed from scratch (in C++) a columnar database to ingest and index in real-time network monitoring flow data from large telcos. Novel compression and indexing algorithms guarantee compression ratios on par with bzip2 and interactive response times (3 patents granted).
- Research outcome: 9 journal and conference papers and 4 patents granted.

Endace *www.endace.com*

SOFTWARE ENGINEER - NETWORK SPECIALIST *August 2008 - February 2009*

- Designed and implemented software for carrier-grade appliances used by 3 of the top 5 telcos in the United States and 2 of the largest exchanges in the world (in the *press*). Worked at multiple level of the stack: operating system, drivers, user space.
- Lead developer of Endace's Wireshark codebase and contributed upstream.

Professional experience (Freelancing)

NEC Network Laboratories *neclab.eu*

SOFTWARE ENGINEER - NETWORK SPECIALIST *June, 2007 - March 2008*

- Delivered a C/C++ carrier-grade software to monitor the quality of VoIP calls in large Telcos using low-cost commodity hardware. To achieve performance on a budget we extended the Linux Kernel with a novel packet-filtering and analysis framework.
- Research outcome: 1 conference paper.

Mutina Technology (now Empirix) *www.empirix.com*

SOFTWARE ENGINEER - NETWORK SPECIALIST *January, 2007 - May, 2007*

- Designed and implemented a C++ library for analyzing network traffic using passive network monitoring technologies. The library provides network performance metrics (e.g. network and application response times) and labels malicious or anomalous traffic (e.g. fragmented ICMP packets, overlapped IP fragments).

Agilent Technologies *www.agilent.com*

SOFTWARE ENGINEER - NETWORK SPECIALIST *January, 2006 - June, 2006*

- With a team of hardware, firmware and software engineers we designed and implemented a network monitoring solution based on custom-built and yet inexpensive FPGA-based SFPs (Small form-factor pluggable transceivers) introducing monitoring capabilities. In 2012 the research prototype turned into the Packet-Portal (*1*, *2*) product from JDSU (*http://www.jdsu.com*).

Education

Ph.D. Electrical Engineering *Zurich, Switzerland*

SWISS FEDERAL INSTITUTE OF TECHNOLOGY (ETH) *Jan. 2010 - May. 2012*

- Thesis: *"High-speed indexing and archival of network measurements data."* Supervisor: Prof. Bernhard Plattner
- The dissertation describes the technologies behind a full commercial system developed while working at IBM Research, Zurich. (3 granted patents)

MSc, Computer Science (top marks with honors) *Pisa, Italy*

UNIVERSITY OF PISA

- Thesis: *"Enterprise Voice-over-IP Traffic Monitoring."* Supervisors: Prof. Luca Deri and Prof. Marco Danelutto
- The dissertation describes an optimized network traffic analyzer developed for NEC Research Laboratories, Heidelberg.

BSc, Computer Science (top marks with honors) *Pisa, Italy*

UNIVERSITY OF PISA

Honors & Awards

2023 **IBM**, *Second Patent Plateau*
2019 **IBM**, *Outstanding Technical Achievement Award*
2017 **IBM**, *Research Division Award*
2015 **IBM**, *Outstanding Technical Achievement Award*
2013 **Red Hat**, *Award for demostrating commitment to Red Hat's Values and Culture*
2011 **IBM**, *Eminence and Excellence Award* for the contributions to the SONAS product
2010 **IBM**, *First Patent Application Award*

Skills

Languages	fluent in C, familiar with C++, Python, Java
Operating Systems	GNU/Linux (kernel development experience), Kubernetes, OSX
Machine learning	Torch, Tensorflow, ONNX, Huggingface ecosystem
Technologies	Kafka, Redis, ElasticSearch, ntop, wireshark, Intel VTune

Patents

A detailed list can be found in my *website* (recent applications might not be searchable yet)

ML/NLP	3 granted, 6 pending
Indexing/Compression	4 granted
Packet processing	2 granted

Open-source

Contributed to the Linux Kernel, openvswitch, ntop, PF_RING, Wireshark.

Publications (Selected)

Machine learning and NLP

pNLP-Mixer: an Efficient all-MLP Architecture for Language
F. Fusco, D. Pascual, P. Staar, D. Antognini, *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*

Extracting Text Representations for Terms and Phrases in Technical Domains
F. Fusco, D. Antognini, *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*

Unsupervised Term Extraction for Highly Technical Domains
F. Fusco, P. Staar, D. Antognini, *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*

RecoNet: An Interpretable Neural Architecture for Recommender Systems
F. Fusco, M. Vlachos, V. Vasiliadis, K. Wardatzky, J.Schneider, *Proc. of the 28th Int. Joint Conf. on Artificial Intelligence (IJCAI 2019)*

On Improving Co-Cluster Quality with Application to Recommender Systems
M. Vlachos, F. Fusco, H. Mavroforakis, et.al, *Proc. of the 23rd Int. Conf. on Information and Knowledge Management (CIKM 2014)*

Indexing and compression

Indexing million of packets per second using GPUs
F. Fusco, M. Vlachos, X. Dimitropoulos, L. Deri, *Proc. of the 13th ACM SIGCOMM Internet Measurement Conference (IMC 2013)*

RasterZip: Compressing Streaming Network Monitoring Data with Support for Partial Decompression
F. Fusco, M. Vlachos, X. Dimitropoulos, *Proc. of the 12th ACM SIGCOMM Internet Measurement Conference (IMC 2012)*

tsdb: A Compressed Database For Time Series
L. Deri, S. Mainardi, F. Fusco, *Proc. of 2012 Traffic Monitoring and Analysis workshop (TMA 2012)*

pcapIndex: An Index for Network Packet Traces with Legacy Compatibility
F. Fusco, X. Dimitropoulos, M. Vlachos, L. Deri, *Computer Communication Review (CCR), Vol. 42, No. 1, Jan 2012*

Real-time creation of bitmap indexes on streaming network data
F. Fusco, M. Vlachos, M. Stoecklin, *VLDB Journal, Vol. 21, No. 3, Jun 2012*

NET-FLi: On-the-fly Compression, Archiving and Indexing of Streaming Network Traffic
F. Fusco, M. Stoecklin, M. Vlachos, *Proc. of the 36th Int. Conference on Very Large Databases (VLDB 2010)*

High-speed packet processing

vPF_RING: Towards Wire-Speed Network Monitoring Using Virtual Machines
A. Cardigliano, L. Deri, J. Gasparakis, F. Fusco, *Proc. of the 11th ACM SIGCOMM Internet Measurement Conference (IMC 2011)*

High-speed Network Traffic Analysis with Commodity Multi-core Systems
F. Fusco and L. Deri, *Proc. of the 10th ACM SIGCOMM Internet Measurement Conference (IMC 2010)*

Wire-Speed Hardware-Assisted Traffic Filtering with Mainstream Network Adapters
L. Deri, J. Gasparakis, P. Waskiewicz Jr, F. Fusco, *Proc. of Network Embedded Management and Applications (NEMA 2010)*

Enabling High-Speed and Extensible Real-Time Communications Monitoring
F. Fusco, F. Huici, L. Deri, S. Niccolini, T. Ewald, *Proc. of the 11th Int. Symposium on Integrated Network Management (IM 2009)*