Stefania Victoria Costache

London, United Kingdom ▪ +447469472926
stefania.costache@gmail.com
linkedin.com/in/scostache

I am a computer engineer with a PhD in cloud compute resource optimization. I have a comprehensive understanding of distributed algorithms and large scale systems and I am also familiar with statistical modeling, machine learning, and data mining concepts. I have a proven track of record in making significant and self-directed contributions to large and challenging projects. I am a strong supporter of Open Source technologies and Agile and DevOps methodologies.

Additional languages include French and Romanian.

Technical Proficiencies

Programming Languages: C, Java, Golang, Python. *Familiar* with C++, Scala, MySQL, Shell, Jupyter Notebook.

Cloud Platform: Kubernetes, Mesos, Docker, OpenTelemetry/OpenTracing; *Familiar* with IBM Cloud, GCP, AWS, OpenStack

Data analytics: NoSQL Datastores/Redis, Timeseries Datastores/InfluxDB, Flink, Spark, Tensorflow, ElasticSearch, Kafka

Professional Experience

J.P. Morgan, London, UK

Software Engineer/Vice President (September 2019 - current)

Building a serverless compute distribution platform for financial risk modelling algorithms

- Contributing in terms of design, implementation and maintenance of the API for compute distribution policies, autoscaling, platform fault tolerance and telemetry using metrics and distributed tracing
- Used data analytics (designed and deployed data pipelines with distributed tracing and different data collection, statistics and visualization tools) to provide observability across the stack for risk calculations on-premise, prediction and optimization of compute costs and give stakeholders insight in aspects of product utilization.

IBM (T.J. Watson Research Center), New York, NY

Research Staff (January 2016 – November 2018)

Conducted research on large scale datacenter resource management. Contributed deliverables to internal MVP projects, co-authored several patents and mentored two research interns. Research topics included:

- Solutions that use dynamic priority and preemption to manage resources for serverless data analytics and deep learning workloads.
- Machine Learning investigations to estimate mentioned workload performance and scheduling requirements.
- Experimental study of scheduling in container cloud frameworks; benchmarked scalability and performance.

Chalmers University of Technology, Sweden

Postdoctoral Researcher (January 2016 – December 2016)

Utilized an intelligent vehicular systems (IoT) benchmark to investigate the performance of two well-known stream processing frameworks, Spark and Flink.

- Conducted an experimental case study that outlined the advantages and limitations of the two technology stacks.
- Published a poster showing the results at a specific international symposium.

continued…

Vrije University Amsterdam, The Netherlands

Postdoctoral Researcher (February 2014 – August 2015)

Co-advised 3 master students and motivated them to successfully complete their projects by helping them brainstorm and draft research papers. Worked on several projects, including (1) the analysis of the scalability of cloud stacks, (2) the development of algorithms for profiling energy consumption; and (3) elasticity of an in-memory datastore for scientific data-intensive workflows.

- Developed an online scheduler for scientific data-intensive workflows that minimizes resource waste by learning task resource demand and adapting the number of co-located tasks.
- Prototyped algorithms and implemented a multi-cloud provisioning policy simulator that revealed hosting cost reductions for PaaS providers.
- Co-authored several conference and journal papers.
- Strengthened student research capabilities by introducing them to different research papers and the methodology of research brainstorm during an advanced master course.

INRIA Rennes-Bretagne Atlantique, France

Research Engineer (September 2013 – January 2014)

Conducted an experimental case study for an European-funded cloud middleware which interfaced with multiple infrastructure clouds and showed that typical application deployments have low performance overheads while also decreasing configuration overheads for users.

EDF R&D/INRIA Rennes-Bretagne Atlantique, France

Research Engineer (May 2010 – July 2013)

Prioritized and scaled application resources by using spot market and virtual economy concepts to design a resource management framework for running HPC applications, such as batch MPI and task farming, on a private cloud.

- Maximized infrastructure utilization by proposing SLO-driven scaling policies for vertical and horizontal resource allocation; this proposal allowed more users to run applications simultaneously.
- Reduced the number of virtual machine migrations below a threshold and decreased resource management performance overheads by developing an incremental placement algorithm for virtual machines.
- Recognized by the French academic community for submitting the best paper on Cloud Computing at the 2014 annual ComPAS meeting.

Internships

INRIA Rennes-Bretagne Atlantique, France, Intern in the Myriads Group, May 2009 – October 2009

University of Groningen, The Netherlands, Intern in the Molecular Dynamics Group, February 2009 – March 2009 and March 2008 – July 2008

Ixia, Bucharest, Romania, Software engineering intern, July 2016 - September 2016

Education

Ph.D., Computer Science, 2013

University of Rennes 1, France

PhD Thesis title: “Market-based autonomous resource and application management in the cloud”
mention "with highest honors"

Master of Computer Science, Automatic Control and Computers Faculty, 2010

University "Politehnica" of Bucharest, Bucharest, Romania

Specialization "Advanced Systems and Applications"
MS Thesis title: "Towards highly available and self-healing grid services"

Bachelor of Engineering, 2008
University "Politehnica" of Bucharest, Bucharest, Romania
Specialization "System and Computer Engineering"
Thesis title: "Multiscaling algorithms for molecular dynamics simulations with GROMACS"

Publications

Selected publications (a complete list is found on Google Scholar):

1. Performance biased scheduler extender. C. Wang, S. V. Costache, A.S. Youssef, A. Kanso, T. Suk, A. N. Tantawi, US Patent App. 16/592,078, 2021
2. Reward-based admission controller for resource requests in the cloud. C. Wang, A. Kanso, S.V. Costache, A.S. Youssef, M. Steinder. US Patent App. 16/204,108, 2020
3. Runtime estimation for machine learning tasks. P. Dube, G. Joshi, P.A. Nagpurkar, S.V. Costache, D. J. Arroyo, Z.N. Sura. US Patent App. 15/901,430, 2019
4. Resource Management in Cloud Platform as a Service Systems: Analysis and Opportunities. S. Costache, D. Dib, N. Parlavantzas, C. Morin. Journal of Systems and Software, Elsevier, 2017
5. MemEFS: A network-aware elastic in-memory runtime distributed file system. A. Uta, O. Danner, C. van der Weegen, A.M. Oprescu, A. Sandu, S. Costache, T. Kielmann. Future Generation Computer Systems, Elsevier, 2017.
6. Market-based autonomous resource and application management in private clouds. S. Costache, S. Kortas, C. Morin, N. Parlavantzas. Journal of Parallel and Distributed Computing, 2016.
7. E-BaTS: Energy-Aware Scheduling for Bag-of-Task Applications in HPC Clusters. A. Vintila-Filip, A.M. Oprescu, S. Costache, T. Kielmann. Parallel Processing Letters, 2015.
8. MemEFS: an elastic in-memory runtime file system for e-science applications. A. Uta, A. Sandu, S. Costache, T. Kielmann. e-Science, 2015.
9. Merkat: A Market-based SLO-driven Cloud Platform. S.V. Costache; N. Parlavantzas; C. Morin; S. Kortas. CloudCom, 2013.
10. Semias: Self-Healing Active Replication on Top of a Structured Peer-to-Peer Overlay. S. Costache, T. Ropars, C. Morin, SRDS, 2010.