

Predictable and Energy-efficient Execution of Distributed Cloud Services: A 2-decades Research Overview

Tommaso Cucinotta
tommaso.cucinotta@santannapisa.it
Scuola Superiore Sant'Anna
Pisa, Italy

Abstract

Information and Communications Technologies have been undergoing a relentless evolution, with the widespread availability of affordable high-speed Internet connections, that caused a major paradigm shift in software development and engineering towards distributed computing. This led to the raise of Cloud Computing [2] as the de-facto standard for managing distributed infrastructures on top of which an ever-increasing number of on-line services are made available, impacting our everyday lives.

Cloud infrastructures have become increasingly attractive for a wide range of modern and future distributed computing scenarios, including those involving the deployment of time-sensitive applications [1], where the ability to provide end-to-end performance and availability guarantees for the hosted applications, in conjunction with resource and energy efficiency, becomes paramount. This is relevant for several application domains: interactive multimedia, virtual and augmented reality and on-line gaming; offloading of advanced functionality in robotics, factory automation, Industry 4.0 and automotive; real-time packet processing and software-defined telecommunications; and others. In these scenarios, it is important to provide performance guarantees to end-users, despite temporal interference and noisy neighbor effects due to resource sharing and multi-tenancy, while tolerating failures at various levels. Among the key challenges a provider has to tackle in this context, there are key design and architectural decisions related to the necessary trade-offs between performance, costs, and energy efficiency.

In this keynote talk, I provide an overview of the research projects I've been overseeing on these topics, spanning nearly 2 decades of research carried out jointly with international academic and industrial partners, highlighting what challenges have been tackled and what mechanisms have been realized. These include our relentless efforts to develop and contribute a real-time CPU scheduler for the Linux kernel to support deadline-driven scheduling of real-time embedded and virtualized workloads, since the early-stage developments within the IRMOS EU project [3, 6, 7], to the integration of SCHED_DEADLINE within the mainline Linux kernel [10, 11], to its hierarchical extension [4] and its proposed integration within the OpenStack [5] and Kubernetes [8, 9] Cloud orchestration platforms.

The talk will conclude with an overview of the challenges that still lie ahead in this area for future research.

CCS Concepts

• **Computer systems organization** → **Real-time systems**; • **Software and its engineering** → **Scheduling**; **Cloud computing**; **Real-time systems software**; **Virtual machines**.

Keywords

Real-Time Systems, Resource Management and Optimization, Cloud Computing, Virtualization, Containers, Linux Kernel

ACM Reference Format:

Tommaso Cucinotta. 2026. Predictable and Energy-efficient Execution of Distributed Cloud Services: A 2-decades Research Overview. In *Companion of the 17th ACM/SPEC International Conference on Performance Engineering (ICPE Companion '26)*, May 04–08, 2026, Florence, Italy. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3777911.3800622>

Biography



Tommaso Cucinotta is Associate Professor at the Real-Time Systems Lab (RETIS) of Sant'Anna School of Advanced Studies in Pisa, Italy. He has a MSc in Computer Engineering from University of Pisa, and a PhD in Computer Engineering from Sant'Anna. His research interests include operating systems, predictable execution and adaptive CPU scheduling for real-time software systems, with reference to a wide variety of platforms, ranging from embedded systems to infrastructures for Cloud and Distributed Computing and Network Function Virtualization (NFV). He has been Member of Technical Staff (MTS) in Bell Labs in Dublin (Ireland), investigating on security and real-time performance of Cloud and NFV services. He has been a senior software development engineer in Amazon Web Services (AWS) in Dublin (Ireland), where he worked on improving the performance and scalability of the DynamoDB real-time NoSQL data-store. Since 2016, he is Associate Professor at Scuola Sant'Anna and coordinator of the Cyber-Physical Research Area since 2019. Tommaso Cucinotta has also brought a number of funded research projects to the RETIS lab, notably EU projects, and international collaborations with private companies, including some big-tech ones. Tommaso Cucinotta coauthored roughly 150 international peer-reviewed scientific publications on topics in his areas of interest, as well as 11 granted and 20 filed patents. He is in the technical program committee of a number of international conferences related to his research topics, and he also performs regularly peer-reviewing for papers submitted to prestigious international journals. He is Associate Editor for the IEEE Transactions on Services Computing and IEEE Transactions on Cloud Computing journals.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

ICPE Companion '26, Florence, Italy

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2326-1/2026/05

<https://doi.org/10.1145/3777911.3800622>

Acknowledgments

This talk summarizes research that has been carried out jointly with several students, colleagues and collaborators, I've been so lucky to have an opportunity to work with, throughout the last two decades. Among these, special thanks are due particularly to Fabio Checconi, Dario Faggioli, Luca Marzario, Antonio Mancina, Juri Lelli, Michael Trimarchi, Alessio Balsini, Daniel Bristot De Oliveira, Remo Andreoli, Stefano Fiori, Luca Abeni, Giuseppe Lipari, Luigi Palopoli, Mauro Marinoni, Manuel Stein, Thomas Voith, Karsten Oberle, Volker Hilt, Carlo Vitucci, Morten Rasmussen, Joel Fernandes, Patrick Bellasi, Harald Gustafsson, Fredrik A. Svensson, Peter Zijlstra, Thomas Gleixner, and too many others to mention them all.

References

- [1] Remo Andreoli, Raquel Mini, Per Skarin, Harald Gustafsson, Janos Harmatos, Luca Abeni, and Tommaso Cucinotta. 2025. A Multi-Domain Survey on Time-Criticality in Cloud Computing. *IEEE Transactions on Services Computing* 18, 2 (2025), 1–19.
- [2] Rajkumar Buyya, Christian Vecchiola, and S. Thamarai Selvi. 2013. Chapter 4 – Cloud Computing Architecture. In *Mastering Cloud Computing*, Rajkumar Buyya, Christian Vecchiola, and S. Thamarai Selvi (Eds.). Morgan Kaufmann, Boston, 3–27. <https://doi.org/10.1016/B978-0-12-411454-8.00001-2>
- [3] Fabio Checconi, Tommaso Cucinotta, Dario Faggioli, and Giuseppe Lipari. 2009. Hierarchical Multiprocessor CPU Reservations for the Linux Kernel. In *Proceedings of the 5th International Workshop on Operating Systems Platforms for Embedded Real-Time Applications (OSPRT)*. Politécnico do Porto, Dublin, Ireland, 15–22.
- [4] Tommaso Cucinotta, Luca Abeni, Mauro Marinoni, Alessio Balsini, and Carlo Vitucci. 2019. Reducing Temporal Interference in Private Clouds through Real-Time Containers. In *Proceedings of the IEEE International Conference on Edge Computing (EDGE)*. IEEE, Milan, Italy, 124–131. <https://doi.org/10.1109/EDGE.2019.00036>
- [5] Tommaso Cucinotta, Luca Abeni, Mauro Marinoni, Riccardo Mancini, and Carlo Vitucci. 2023. Strong Temporal Isolation Among Containers in OpenStack for NFV Services. *IEEE Transactions on Cloud Computing* 11, 1 (2023), 763–778. <https://doi.org/10.1109/TCC.2021.3116183>
- [6] Tommaso Cucinotta and Fabio Checconi. 2010. The IRMOS realtime scheduler. <https://lwn.net/Articles/398470/>
- [7] Tommaso Cucinotta, Fabio Checconi, George Kousiouris, Dimosthenis Kyriazis, Theodora Varvarigou, Alessandro Mazzetti, Zlatko Zlatev, Juri Papay, Michael Boniface, Soeren Berger, Dominik Lamp, Thomas Voith, and Manuel Stein. 2010. Virtualised e-Learning with Real-time Guarantees on the IRMOS platform. In *IEEE Int. Conf. on Service-Oriented Comp. and Appl.* IEEE, Perth, Australia, 1–8. <https://doi.org/10.1109/SOCA.2010.5707166>
- [8] Stefano Fiori, Luca Abeni, and Tommaso Cucinotta. 2022. RT-kubernetes: containerized real-time cloud computing. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (SAC'22)*. ACM, Brno, Czech Republic, 36–39. <https://doi.org/10.1145/3477314.3507216>
- [9] Harald Gustafsson, Fredrik Svensson, Raquel Mini, Luca Abeni, Remo Andreoli, and Tommaso Cucinotta. 2025. RTilience: Fault-Tolerant Time-Critical Kubernetes. *IEEE Transactions on Services Computing* 8, 16 (2025), 1–14.
- [10] Juri Lelli, Dario Faggioli, Tommaso Cucinotta, and Giuseppe Lipari. 2012. An experimental comparison of different real-time schedulers on multicore systems. *Journal of Systems and Software* 85, 10 (2012), 2405–2416. <https://doi.org/10.1016/j.jss.2012.05.048> Automated Software Evolution.
- [11] Juri Lelli, Giuseppe Lipari, Dario Faggioli, and Tommaso Cucinotta. 2011. An efficient and scalable implementation of global EDF in Linux. In *Proceedings of the 7th International Workshop on Operating Systems Platforms for Embedded Real-Time Applications (OSPRT)*. The George Washington University, Porto, Portugal, 6–15.