Trends, drivers, and strategic directions for trustworthy edge computing in industrial applications

Authors: James Gross, Martin Törngren, György Dán, David Broman, Erik Herzog, Iolanda Leite, Raksha Ramakrishna, Rebecca Stower, and Haydn Thompson

Abstract: TECoSA - a university-based research center in collaboration with industry - was established early in 2020, focusing on Trustworthy Edge Computing Systems and Applications. This article summarizes and assesses the current trends and drivers regarding edge computing. In our analysis, edge computing provided by mobile network operators will be the initial dominating form of this new computing paradigm for the coming decade. These insights form the basis for the research agenda of the TECoSA center, highlighting more advanced use cases, including AR/VR/Cognitive Assistance, cyber-physical systems, and distributed machine learning. The article further elaborates on the identified strategic directions given these trends, including an emphasis on testbeds and collaborative multidisciplinary research.

Keywords: Edge computing, cyber-physical systems, trustworthiness, systems engineering, innovation eco-systems

Introduction

Several trends and drivers interact in the ongoing digitalization shift including edge computing, connectivity, artificial intelligence, and big data loops, where field data are gathered to continuously update software systems. This transformation offers unprecedented innovation and product development opportunities, and also act as enablers for industrial companies to meet their targets for sustainable development goals. The need to address all dimensions of sustainability is highlighted by the recent European Commission initiative on Industry 5.0, emphasizing that previous efforts such as Industry 4.0 have predominantly focused on productivity, (EC Industry 5.0, 2022). A concrete example of what CPS can do for sustainability are the "tools" available to facilitate circularity, e.g., with traceability and predictive capabilities to support decisions regarding maintenance and recycling. However, the digital transformation also increases system complexity and introduces challenges of a socio-technical nature, such as risks related to technical systems acting in open environments including ethical considerations related to fairness and personal integrity, INCOSE (2021), Törngren (2021). Specifically, our future societies will be dependent on increasingly sophisticated infrastructures where **edge computing** will act as a new tier, complementing the cloud and embedded systems. TECoSA, a research center on trustworthy edge computing systems and applications, was formed and initiated in 2020 to address the corresponding key challenges, TECoSA (2022), Törngren et al. (2021). The center brings together multiple research teams at KTH Royal Institute of Technology and (currently) 15 industrial partners, spanning several industrial domains. The discussions among the center partners form the basis for the results presented in this paper.

TECoSA is active in industrial digitalization with a focus on edge computing systems. The aim is to provide methods, tools, and theories for building trustworthy systems relying on edge computing. The emphasis during the initial phase of the center—in the context of trustworthiness—has been on safety, cyber-security, and predictability (see Figure 1). Trustworthiness has traditionally been associated with human-machine interactions and security, referring to how we (humans) perceive trust in relation to services and machines. Trustworthiness has evolved to become an umbrella term encompassing the concept of dependability, associated with properties like reliability, availability, maintainability, safety and security, and properties associated with artificial intelligence such as transparency, explainability and fairness, AI HLEG (2021).

The main purpose of this article is to initially summarize and assess the current trends and drivers regarding edge computing. These insights form the basis for the research agenda of the TECoSA center. As a second purpose, the article elaborates and discusses the identified strategic directions given these trends.

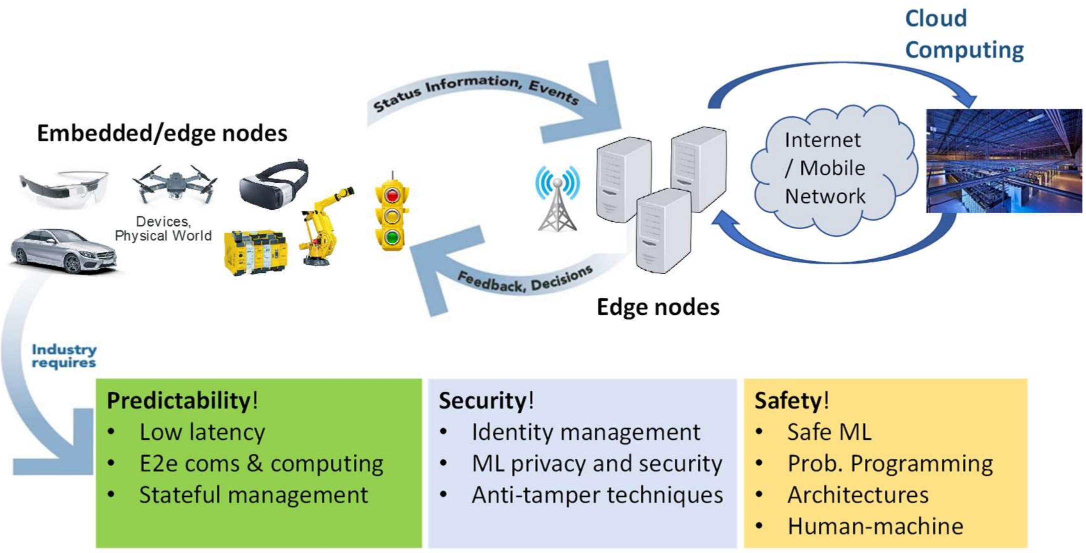

Figure 1: Edge computing as a new tier complementing embedded systems (device edge) and the cloud, illustrating initial trustworthiness properties and challenges addressed by the TECoSA center

Edge Computing State-of-the-Art

Edge computing is perhaps best understood in its contrast to cloud computing. In the 2000s, the first-generation internet architecture was dominated by the client-server approach. Most of the clients were desktop PCs on private or corporate premises, connected via the Internet to web servers. A private or corporate entity intending to offer information or services on the Internet had to acquire server hardware and software, install and maintain it on corresponding premises, and set up a matching Internet connection. By 2010, this division changed dramatically. On the one hand, an increasing fraction of the clients were mobile devices, connecting through mobile networks like 3.5G and the upcoming 4G (LTE) to the Internet. On the other hand, web service offerings moved more and more to cloud providers, where very large pools of server hardware were brought together, allowing a scalable and efficient operation of web services from an installation, maintenance, and connectivity point of view. Web service operators moved from hosting and maintaining servers (together with the content) locally on premise, to only curating content while renting the hardware and software for the web service from cloud providers. As a result, cloud computing centers of corresponding providers often ended up in locations where physical space, energy supply, and backbone connectivity were cheap, resulting in rather remote locations. By and large, this is the dominating service model of the Internet as of today.

In this context, edge computing is primarily defined as the offering of compute services in “closer physical proximity” to clients compared to cloud computing, i.e., offering compute services towards the “edge” of the Internet / wide-area networks. Given the dominant presence of 4G and 5G mobile networks as primary access networks of a majority of clients in today’s Internet, edge computing is hence realized by placing corresponding compute resource, either within the mobile network core, or even within a radio access network depending on the preferred proximity. In this line of thinking, proximity is traded with scale and cost: The higher the desired proximity of edge compute resources to the mobile clients , the more physical locations for cloudlet placements will be required, leading also typically to less computational resources available per edge compute location.

Visions associated with edge computing have in various academic/industrial communities been given different names, including for instance multi-access edge computing (MEC) (related to telecommunications and 5G, earlier referred to as mobile edge computing), (Abbas et al., 2018), fog computing (with localized computations through communication devices such as routers and gateways in collaboration with the cloud), (Bonomi et al., 2012), and cloudlets (small scale localized data centers), (Satyanarayanan, 2017). In the current discourse, edge computing has moreover been associated with either locality, computing technologies, or both (Varghese et al., 2021). While the many projections for edge computing may appear as confusing, this situation is not surprising since we are in the early stages of edge computing with an ongoing market positioning.

Our analysis from a commercial point-of-view is that edge computing provided by mobile network operators will be the initial dominating form of this new computing paradigm for the upcoming decade. Beyond that, new concepts might arise that exploit more a continuum of available compute points along the route from mobile clients to cloud centers, see e.g., Duranton et al. (2021). *In the following, we therefore refer to edge computing as the provisioning of additional computing resources through mobile networks.* Edge computing could be introduced to decrease hardware cost in mobile devices, such as industry robots and civilian or military surveillance systems, while meeting latency and predictability demands. In this sense, edge computing adds computational resources that complement existing capabilities of devices (embedded systems) and the cloud, belonging to a tier of a digitalized infrastructure. For a presentation of more detailed use cases, see the following discussion below.

Edge computing has arguably been introduced roughly twenty years ago under the synonym "cyber foraging" (Balan et al. 2002). Since then, a set of various arguments have been brought up highlighting potential benefits of edge computing:

- The original cyber-foraging research had been motivated by the desire to **improve energy-efficiency** if compute-intensive jobs could be offloaded from battery-powered mobile clients to stationary, but close-by cloudlets, also decreasing network-wide energy consumption. Either code or input data is offloaded through a mobile network to cloudlets, with the result of the computation being sent back to the client. Cloud computing is seen in this context as having too long latency and being too unreliable, necessitating edge computing.
- A second argument, related to the above, can be made about the relative distance of cloud computing centers and therefore in particular a much **lower access delay** in case of edge computing. For compute tasks that are either too complex for mobile clients, or that require input from multiple mobile clients, while being latency-sensitive, edge computing provides a clear advantage in providing lower round-trip delays. This case is for instance often made in the context of augmented or extended reality applications.
- Edge computing can also drastically **reduce the bandwidth** required for certain analysis services that run in the cloud. In this case, cloudlets are used as primary processing units, for instance with respect to video analytics in detecting certain events or states in the video stream. Instead of conveying the entire stream to a cloud center, leading to a large bandwidth requirement as more and more end points are included in the service, only indices of the video frames and of the detected objects are provided upstream to the cloud center. The corresponding video frames are nevertheless stored at the edge and can be retrieved by the cloud center. Similar cases can be made for predictive maintenance, IoT systems, as well distributed machine learning applications.
- Finally, edge computing systems come with **different security and privacy features**. While typical concerns of security and privacy with respect to cloud compute centers do not carry over to edge computing, new aspects such as physical access and manipulation become more relevant in the case of edge computing. Related to this shift towards more "local" aspects of security and privacy are also advantageous of edge computing with respect to **regulatory frameworks**. Due to the geographical proximity of deployed cloudlets and corresponding

clients, edge computing offerings might guarantee the manipulation and storage of data to happen within a certain regulatory framework, which a general-purpose cloud provider might not be able or willing to guarantee (in contrast to sovereign cloud offerings).

From these diverse drivers and advantages discussed in the academic/industrial community over the last ten years, for the first wave of commercial edge computing offerings foreseeable today, the regulatory and the bandwidth saving aspects appear to be the main drivers. With respect to B2B customers, edge computing offerings of mobile network providers, referred to as Telco edge, as well as cloud providers, referred to as regional cloud, will offer guarantees with respect to the compute and storage location, and therefore the regulatory conditions under which data is manipulated and/or stored. In addition, hybrid edge-cloud solutions are emerging that push the bulk of the processing to edge cloudlets while integrating the results of local cloudlet-based compute with cloud services. It is important to note that in both cases "best effort" service level agreements (SLAs) between service provider and customer are sufficient for successful commercialization. Beyond these B2B offerings, in the B2C space a prominent commercialization case for edge computing appears to be online multi-player gaming, where depending on the location of the players and on the placement of the game backend process, significantly higher quality of experience can be achieved. Still, corresponding offerings in the gaming domain will be run under best effort SLAs.

Beyond "Best effort"

More advanced use cases exist that could benefit from edge computing but where different challenges exist as of today, including both technical/scientific as well as related to business models. The commercial viability of these opportunities thus remains uncertain, and the TECoSA center has therefore identified three types of use cases as particularly interesting where more research is needed. These use cases all demand more localized computing power, providing incentives for edge computing. The use cases are also relevant in several application domains, driving setups in which a digitalized edge computing-based infrastructure promises to provide added value (see Figure 2). In manufacturing, for example, many ongoing field tests involve the use of private 5G networks and edge computing, representing such digitalized infrastructures. We first elaborate these use cases and then discuss approaches to address them.

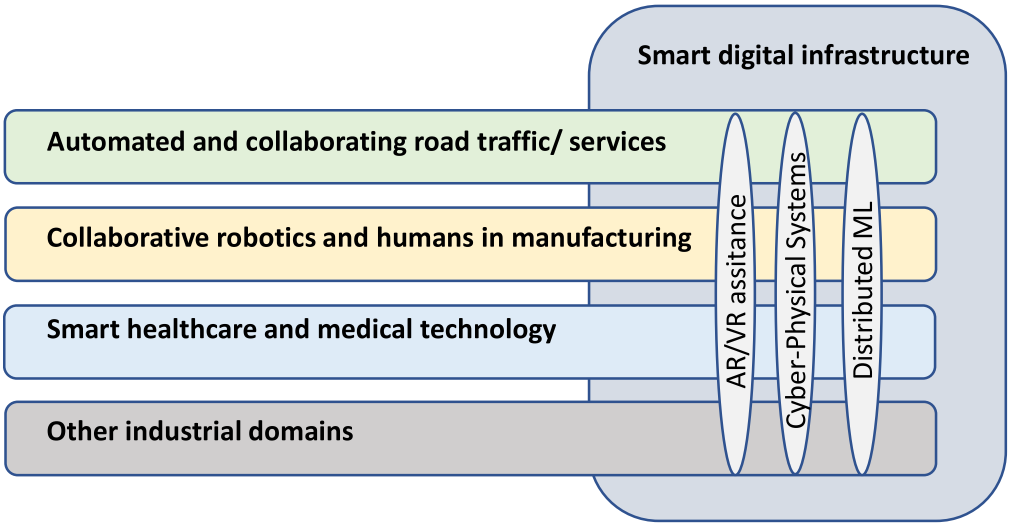
Smart digital infrastructure
Automated and collaborating road traffic/ services
Collaborative robotics and humans in manufacturing
Smart healthcare and medical technology
Other industrial domains
AR/VR assitance
Cyber-Physical Systems
Distributed ML

Figure 2: Various application domains, use-cases of cross-domain relevance, and interactions with a digital infrastructure (providing edge computing, communication and other capabilities such as positioning)

- Use case 1: Mobile AR/VR/Cognitive Assistance: The first use case concerns advantages of future edge computing deployments in human-in-the-loop applications like virtual reality (VR) and augmented reality (AR). These are closed-loop systems where different "status" information is conveyed upstream to the point-of-computation (i.e., the cloudlet). At the backend, the provided status information is used for generating feedback, which is generated and transmitted back to the application client. AR and VR applications are generally characterized by (1) high data rate requirements upstream and/or downstream, (2) complex backend processing taking place at the cloudlet, and (3) quality-of-experience (QoE) of the application being directly related to the responsiveness of the entire loop (upstream communication, compute, and downstream communication). Subtle differences exist with respect to the workloads and QoE requirements for AR systems versus VR systems, where VR systems require higher bandwidths in the downlink and, generally, speaking the latency requirements are also higher due to the level of immersion. The specific challenges with respect to both application types relate to:

1. Efficient application support: Due to the interplay between communication and compute elements over the offloading loop, many trade-offs exist with respect to dynamically managing end-to-end delays at runtime. These trade-offs are largely unexplored, particularly, in relation to quality-of-experience implications in the short- and long-term. Managing end-to-end delays with respect to QoE over a heterogeneous set of active AR/VR applications is a further challenge, as is the question of optimal placement of the compute backend or an efficient and reliable mobility support for such applications. To a large extent, the efficient support of such applications hinges also on the degree of control the application will be able to execute over the mobile network. In the past, mobile network systems have offered only very limited APIs (application programming interfaces) as QoE requirements for voice, video or web applications have been largely similar and hence easy to manage. However, for AR or VR applications, more complex trade-offs exist which are likely to be only known to the application at runtime. Hence, a more powerful API for resource control enables a significantly more efficient operation.

2. Scalable life cycle support of applications and end system acceptance: While several SDKs exist for AR and for VR systems, the process from devising a new application over programming, deployment, and updates, is highly complex and requires deep software engineering and platform knowledge. This is in contrast to the corresponding life-cycle support of smart phone apps of various existing ecosystems currently present in the market. From the perspective of the supply side of future AR/VR applications, a significant simplification of the life cycle support is likely to be established over the next years. In part due to the above limitations, AR technology commercialization has been limited. Advanced designs in combination also with a changing sentiment in the group of early adopters might lead over the next years to a break-through of these applications. A scalable provisioning of backend compute capabilities via edge computing paired with a near ubiquitous mobile network access will certainly lift the technological bottlenecks for wide-spread adoption.

- Use case 2: Cyber-physical systems (CPSs): CPSs represent the "integration of computation, networking and physical processes". While CPSs have been around at least since the 1970s with the integration of microprocessors with physical systems, these systems are now seeing unprecedented potential in their capabilities, see e.g., Thompson and Reimann (2018). Representative examples include automated vehicles and future manufacturing systems. In such CPSs, additional sensors, communications, and collaboration can be used to enhance context awareness and planning. The role of edge computing comes in to play to provide the needed computational and analytics support, providing potential for handling large amounts of data for real-time applications, and also supporting CPS collaboration. TECoSA has identified many such applications in domains such as those depicted in Fig. 2, supporting enhanced quality and new functionalities, e.g., through ensuring that the right assembly tools are used for the right parts in a manufacturing process. For CPSs we identify the following challenges:

1. Holistic management of computing and communication resources: Industrial applications come with demanding requirements on real-time (predictable and short enough) latencies, availability and error detection and handling. This requires novel end-to-end resource management capabilities, including exploiting an interplay between applications and infrastructure, and the consideration of energy consumption as a key metric, such that the use of edge computing also minimizes/reduces overall energy consumption of applications.

2. Trustworthy applications based on edge computing: As already introduced, trustworthiness has evolved to become an umbrella term. Given the evolution of CPS, most of the trustworthiness properties will be relevant for future CPS. Incorporating edge computing into future CPS and collaborating CPS, poses new challenges given new failure modes and cyber-security risks (vulnerabilities) of edge computing-based infrastructures and applications. The dependencies and trade-offs between trustworthiness properties require specific attention, especially for open and collaborative CPS with potential conflicts between cyber-security, safety, availability and data sharing. The uncertainty involved in such open further CPS requires run-time risk assessment and handling/adaptation to appropriately balance e.g. safety and availability/performance. Certification and re-certification of (evolving and adapting) edge-based CPS also represents an open challenge.

3. Collaborating systems and scalability: Collaborating systems, often referred to as systems of systems (SoS), lack a central authority responsible for systems integration and where the constituent units evolve independently (e.g., in the domain of roads, both actors such as vehicles and the physical and digital infrastructures of the roads), Maier (1998). This leads to challenges regarding the overall design and responsibilities of such SoS, and also strongly relates to the business model(s) and liability if something goes wrong. The example of "intelligent transport systems" has shown the difficulty of establishing such SoS. We believe that the introduction of 5G and beyond as a digital infrastructure, with its provision for low latency and managed quality of service, may help to create the momentum needed to establish the required models for collaboration.

- Use Case 3: Distributed ML: Machine learning (ML) is widely considered an efficient tool for optimization, prediction, and classification tasks found in various industrial and consumer applications, among others in AR/VR systems (UC1) and CPSs (UC2). The use of ML in these systems could be limited to the application of the pre-trained model for performing a certain task on data received from end devices, referred to as inference. More generally, it could also entail periodic training of the model in order to adapt it to changing environmental conditions. The use of ML for inference usually involves upstream traffic and may involve downstream traffic also if/when the inference leads to decisions that in turn affect devices. Training of a model may involve downstream traffic as well if the updated model is to be distributed to end-user equipment. ML algorithms are often represented as execution graphs and can be found deployed on a variety of devices spanning the edge to cloud continuum. Such distribution of ML primitives enables capabilities previously unattainable in energy and computationally constrained environments. E.g., by placing parts of the execution graph that have hard real-time requirements and low computation complexity on end user equipment, and placing computationally intensive parts in the edge cloud, one can obtain low latency ML algorithms with limited computational resources. At the same time, distributed ML comes with a variety of challenges, in particular:

1. Interoperability: Interfaces for interconnection are needed to enable interoperability between components from different vendors and for making system integration more cost efficient. Considering that the development of ML algorithms is in a fairly early stage, establishing interfaces that will last years or decades is challenging.

2. Systems architecting: Systems architecting aspects and algorithmic issues will become key in ever more complex installations. It is so far unclear how to formulate architectural and design principles for complex, ML-enabled systems in a way that ensures functional and non-functional requirements, and at the same time allows for efficient life-cycle management. A closely related issue is that of sustainability, both in terms of energy consumption and in terms of the environmental footprint of the compute and communications infrastructure needed for ML integration.

3. Robustness and cybersecurity. Robustness to adversarial environments and the lack of privacy guarantees could also hinder the wide scale adoption of ML-enabled systems. ML algorithms have been shown to be vulnerable to adversarial inputs, e.g., minor perturbation of the data, unnoticeable manipulations of algorithm parameters, and trained ML models may also reveal confidential information about the data set used for creating them (Ramakrishna 2022). These issues related to trustworthiness remain to be solved.

The TECoSA center approach to address these challenges: Successful research centers have been reported to exhibit characteristics including that of conducting collaborative multidisciplinary research involving multiple domains, use of testbeds/demonstrators, and having a strong connection to education, (Patterson, 2014). We agree these characteristics are important. TECoSA has placed emphasis on creating a knowledge eco-system with the involved stakeholders and has the goal in the coming period to develop testbeds as experimental and open infrastructures in the areas of automated and connected road traffic, and collaborative robotics. These testbeds will be used to support collaborative research and education. One important aspect of the testbeds is to stimulate the interplay between applications - potentially involving all the types of mentioned use-cases - and digital infrastructures (recall Figure 2). This interplay corresponds to interactions between different research teams and organizations/companies, places the focus on platforms and services (the interfaces between applications and the infrastructures), and moreover have the potential to be used in education and to stimulate open debate on the socio-technical implications.

Conclusions and Future Work (convergence of the use cases etc.)

We have discussed trends and drivers related to edge computing, as a new computing tier overcoming limitations of, and complementing, the cloud and resource constrained embedded systems. The multitude of concepts such as MEC, cloudlets, fog computing, “near/far/nano/enterprise edge”, and “distributed cloud”, (see e.g. Heinen, 2021), while partly confusing, is natural considering that we are in the early stages of edge computing with an ongoing market positioning. In our analysis, edge computing provided by mobile network operators will be the initial dominating form of this new computing paradigm for the coming decade. In this sense, edge computing adds computational resources that complement the existing capabilities of devices (embedded systems) and the cloud, belonging to a tier of a digitalized infrastructure. The first wave of commercial edge computing offerings is likely to be driven by regulatory aspects, bandwidth saving, and soft real-time interactions such as gaming. We highlight that in these cases “best effort” SLAs between service provider and customer are sufficient for successful commercialization.

We have also discussed more advanced use cases including AR/VR/Cognitive Assistance, CPSs and distributed ML, and corresponding challenges that require further research. It is clear that the three presented use cases will in many ways be part of the same system, for example with humans in the loop (e.g., “cobots”—humans and robots collaborating) in the context of CPS, and with data gathering and distributed machine learning taking place in parallel with the other use cases.

In addressing these use cases and challenges, the identified key role of a university led research center is to maintain and grow a knowledge ecosystem to support innovation, research, and education in trustworthy edge-based CPS, and to develop corresponding technological foundations and methodologies.

References

Abbas N., Zhang Y., Taherkordi A., and Skeie T. (2018). *Mobile Edge Computing: A Survey.* IEEE Internet of Things Journal 5, 1 (2018), 450–465.

AI HLEG (2021). *High-Level Expert Group on AI of European Commission. Overview of deliverables from the AI HLEG.* Web page reference: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (accessed 2022-08-17).

Balan R., Flinn J., Satyanarayanan M., Sinnamohideen S., and Yang H. (2002). *The case for cyber foraging*. Proc. 10th workshop on ACM SIGOPS European workshop (EW 10). Association for Computing Machinery, New York, NY, USA, 87–92. https://doi.org/10.1145/1133373.1133390

Bonomi F., Milito R., Zhu J., and Addepalli S. (2012). *Fog Computing and Its Role in the Internet of Things*. Proceedings 1st Edition MCC Workshop on Mobile Cloud Computing (Helsinki, Finland) (MCC '12). ACM, New York, NY, USA, 13–16. https://doi.org/10.1145/2342509.2342513

Duranton M., Malms M. and Ostasz M. (2021). *The continuum of computing*. Hipeac Vision 2021. https://doi.org/10.5281/zenodo.4719341

EC Industry 5.0 (2022). Web page reference: https://research-and-innovation.ec.europa.eu/research-area/industry/industry-50_en (accessed 2022-08-17).

INCOSE (2021). *Systems Engineering Vision 2035.* https://www.incose.org/about-systems-engineering/se-vision-2035

Heijnen A. et al. (2021). *IoT and Edge Computing: opportunities for Europe.* June 2021. Report by the NGIoT project (Next Generation Internet of Things). Retrieved from https://www.ngiot.eu/

Maier M. (1998). *Architecting principles for systems-of-systems.* Systems Engineering Journal. Vol. 1, no. 4, pages 267-284, 1998.

Patterson D. (2014). *How to build a bad research center.* Commun. ACM 57, 3 (March 2014), 33–36. https://doi.org/10.1145/2566969

Satyanarayanan M. (2017). *The Emergence of Edge Computing.* IEEE Computer 50, 1 (2017).

Törngren M. (2021). *Cyber-physical systems have far-reaching implications.* Hipeac Vision 2021. https://doi.org/10.5281/zenodo.4710500

TECoSA, (2022). Web page reference: https://www.tecosa.center.kth.se/ (accessed 2022-08-17).

Ramakrishna R. and Dán G. (2022). *Inferring Class-Label Distribution in Federated Learning.* ACM Workshop on Artificial Intelligence and Security (AISec), Nov. 2022

Ramli R. and Törngren M. (2022). *Towards an Architectural Framework and Method for Realizing Trustworthy Complex Cyber-Physical Systems.* Joint Proceedings of RCIS 2022 Workshops and Research Projects Track, May 17-20, 2022, Barcelona, Spain.

Thompson H. and Reimann M. (2018). *Platforms4CPS Key Outcomes and Recommendations.* https://www. platforms4cps.eu

Törngren M., Thompson H., Herzog E., Inam R., Gross J. and Dán G. (2021). *Industrial Edge-based Cyber-Physical Systems - application needs and concerns for realization*. Proc. of ACM Symp. on Edge Computing Workshop on Trustworthy Edge Computing, Dec. 2021

Varghese B. et al. (2021). *Revisiting the Arguments for Edge Computing Research.* IEEE Internet Computing, doi: 10.1109/MIC.2021.3093924.