Toward Federated Cognitive Digital Twins over the Edge–to-Cloud Continuum

Alessandra Somma
University of Naples Federico II, Italy

Alessio Bucaioni
Mälardalen University, Sweden

Abstract

Digital Twins (DTs) are increasingly adopted to monitor, analyze, and optimize Cyber-Physical Systems (CPSs) by enabling continuous interaction between physical assets and their digital representations. Despite their potential, current DT architectures face significant limitations when applied to distributed environments such as smart cities. In particular, most existing solutions rely on centralized and monolithic designs, which introduce latency, scalability, and resilience issues. Morever, they provide limited support for semantic integration and high-level reasoning, which hinders the effectiveness of DTs-based decision-making.

Recent research has explored Federated Digital Twins (FDTs) to decompose complex systems into multiple interacting twins, improving scalability and modularity. However, existing FDT approaches often lack a clear architectural framework and still concentrate intelligence in cloud-based components, thus limiting local autonomy. In parallel, Cognitive Digital Twins (CDTs) have been proposed to enhance DTs with semantic reasoning and explainability, leveraging Artificial Intelligence and Large Language Models (LLMs). While promising, these approaches are typically centralized and difficult to integrate within distributed architectures.

In this paper, we propose a unified approach that combines federation and cognition within a single architecture, referred to as the Federated Cognitive Digital Twin (FCDT). The proposed architecture distributes intelligence across the edge-to-cloud continuum by introducing local twins deployed close to physical systems and global twins operating at the cloud level. Local twins provide real-time monitoring, analysis, and first-level decision-making, enhanced by lightweight cognitive capabilities; global twins perform system-level reasoning, simulation, and coordination, and leverage more computationally intensive models for explanation and decision support.

By combining distributed autonomy at the edge with global cognitive simulation and reasoning in the cloud, the FCDT architecture addresses both structural and semantic limitations of current DT solutions. It enables scalable and responsive DT systems, improves decision-making capabilities, and simplifies the engineering of complex CPSs in distributed environments.

2012 ACM Subject Classification Software and its engineering → Software design engineering; Computer systems organization → Cloud computing

Keywords and phrases Digital Twins, Cloud Continuum, Edge Computing, Federation, LLMs

Digital Object Identifier 10.4230/OASIcs.CVIT.2016.23

1 Introduction

Digital Twins (DTs) are virtual representations of Cyber-Physical Systems (CPSs), characterized by continuous bidirectional communication between the physical and digital domains [7]. This interaction enables the transfer of real-world data from the physical system to its digital counterpart, while also allowing the DT to generate insights and provide feedback to the physical world [4, 25]. By combining data, models, and analytics, DTs support advanced monitoring, enhance situational awareness, and enable data-driven decision-making across domains such as manufacturing, transportation, and smart cities [2, 23].

Engineering and architecting DT is a complex and demanding activity, and for this reason

it has received growing attention in recent years. This interest has led to the proposal of many DT architectures [7,15], together with important standardization efforts [6,9]. However, when DTs are applied to large-scale and inherently distributed environments such as smart cities, new challenges emerge [23]. In these contexts, heterogeneous devices continuously generate data, multiple subsystems interact with one another, and decisions often need to be taken in a timely and coordinated way. As a result, current DT approaches reveal two major weaknesses: they are still largely based on centralized designs [5, 7, 24], and they provide limited support for semantic integration across heterogeneous subsystems and information sources [8, 24].

Most existing DT architectures address this complexity through centralized solutions. In centralisez designs, the DT is usually implemented as a monolithic system, often deployed in the cloud, and designed as a single unit representing the whole real system. At the same time, data generated by distributed physical assets are transmitted to centralized platforms, where are then stored and processed to feed simulation models, analytics pipelines, and AI-based inference engines. This design has clear advantages: it simplifies DT management, supports large-scale data processing, and enables advanced monitoring and simulation capabilities. However, these benefits come at a cost. The concentration of both the DT logic and the data processing pipeline in centralized infrastructures introduces significant challenges in terms of latency, bandwidth consumption, and resilience [20, 21]. Centralized solutions remain also sensitive to network disruptions and make local autonomy difficult to achieve [17].

For these reasons, DT research is progressively moving beyond monolithic and fully centralized designs toward more distributed solutions. In this direction, the concept of **Federated Digital Twins** (FDTs) has emerged as a promising way to decompose the complexity of large CPSs [26, 27]. The main idea is to model the overall system as a set of autonomous yet interoperable twin instances, each associated with a specific subsystem, while allowing them to cooperate in order to provide a coherent system-level view [16, 27]. In this way, federation addresses an important structural limitation of monolithic DTs, since the complexity of the global system is no longer handled by a single twin, but distributed across multiple coordinated twins.

However, the FDT paradigm is still in an early stage, and many existing proposals remain at a rather high level of abstraction. In particular, they often lack a clear architectural framework to support coordinated intelligence across distributed twins [14, 23, 28]. Moreover, even when federation is adopted, intelligence is frequently still concentrated in cloud-based components. This helps in managing the overall complexity of the DT, but it does not fully solve the problems that motivated the shift away from centralization in the first place, especially latency, resilience, and local autonomy. The result is that currently federation of DTs improves their structural decomposition, but often without yet providing a fully effective distribution of their intelligence [14, 28].

Hence, the shift toward federation helps in addressing the structural complexity of distributed CPSs [13, 27], but it does not fully solve their limited semantic capabilities. As a consequence, DTs are often able to detect or predict events, but they struggle to explain them, relate them across subsystems, or support reasoning about their implications. For example, identifying a delay in a transportation network does not automatically provide an understanding of its causes, how it propagates through the system, or which actions should be taken to mitigate its impact. This gap between data processing and meaningful reasoning limits the effectiveness of DTs as decision-support systems [12].

To address this limitation, recent research has introduced the concept of **Cognitive Digital Twins** (CDTs) [12, 31], which extend traditional DTs with capabilities for semantic

reasoning, knowledge integration, and explainability. The idea is to move from DTs that only process data to DTs that can also interpret it, relate it to domain knowledge, and support human understanding [1, 30]. In this context, advances in Artificial Intelligence (AI), and in particular in Generative AI (GenAI), play a key role. Large Language Models (LLMs) have shown strong capabilities in integrating heterogeneous information, reasoning across domains, and interacting with users through natural language [12, 29, 32].

However, introducing this type of intelligence also raises new challenges. LLMs typically require significant computational resources and are commonly deployed in centralized environments. This creates a tension with the need for distributed DT architectures, especially in scenarios where low latency and local autonomy are required [22]. If not carefully designed, the integration of such models risks reinforcing centralization, which is precisely one of the limitations that federated DTs aim to overcome.

Recent developments suggest that this tension can be mitigated by distributing intelligence across the edge–to-cloud continuum [3, 10, 11, 19]. In this setting, lightweight AI models, including distilled or specialized versions of language models, can be deployed at the edge to support local tasks such as filtering, summarization, and contextual interpretation [18]. At the same time, more computationally intensive models can be deployed in the cloud to provide global reasoning, explanation, and decision support [3, 10].

In this paper, we propose to address these challenges by proposing a unified architecture combining federation and cognition, resulting in the **Federated Cognitive Digital Twin** (FCDT) architecture. In this architecture, local twins provide fast and autonomous decision-making close to the physical system, while global twins support system-level reasoning and coordination. By combining these two levels, the proposed approach maintains responsiveness while enabling advanced analysis and explanation, and at the same time simplifies the engineering of complex DT systems and mitigates latency issues in distributed environments.

2 Related Work

Existing research on engineering and architecting DTs has evolved along multiple directions, addressing different aspects of DTs design, distribution, and intelligence. In this section, we focus on two lines of research, which are relavant to our work, i.e., Federated DTs and Cognitive DTs.

Early DT architectures are predominantly based on centralized designs, where data from distributed assets are collected and processed in cloud environments [5, 7, 15, 24]. FDTs have been introduced to address the structural limitations of monolithic DT architectures in large-scale systems. Instead of representing the entire real system through a single DT instance, FDT approaches decompose it into multiple interacting twins, each associated with a specific subsystem.

Vergara *et al.* [26,27] define FDTs as a distributed set of twin instances that collaborate to support system-level decision-making. Their work emphasizes the role of federation in enabling collaborative simulation and coordination, where different twins exchange information to improve global understanding. Papacharalampopoulos *et al.* [16] analyze federation from a modeling perspective, highlighting limitations in knowledge transfer and interoperability between twin instances. Their work shows that even when multiple twins are connected, inconsistencies and lack of shared semantics can hinder effective collaboration. This reinforces the need for stronger integration mechanisms beyond simple federation.

Model-driven approaches, such as the one proposed by Michael *et al.* [14], investigate how FDTs can be engineered using abstraction and modeling techniques. These approaches

contribute to structuring the development process, but they still operate at a high level and do not explicitly address runtime coordination or distributed decision-making.

More recent efforts, such as [28], explore learning-based coordination across federated twins, while large-scale initiatives like interTwin [13] combine federation with distributed computing infrastructures. These works show the potential of FDTs in complex environments, but they also reveal a common limitation: intelligence is often still centralized, typically in cloud-based components that aggregate and process data from all twins.

As a result, current FDT approaches successfully address structural decomposition and scalability, but only partially address distributed intelligence. In many cases, federation is used to organize the system, but decision-making and reasoning remain centralized, which limits responsiveness and local autonomy.

In parallel, CDTs have been proposed to address the limited reasoning capabilities of traditional DTs. CDTs extend DTs with semantic reasoning, knowledge integration, and explainability, moving beyond purely data-driven analytics [12,31]. Zheng *et al.* [31] introduce the concept of cognitive DTs as systems that integrate data, models, and knowledge to support decision-making. Their work highlights the importance of combining physical models with higher-level reasoning mechanisms, but does not provide a concrete architectural realization.

Early CDTs works such as [1] and [30] propose hybrid DT architectures that combine data-driven models with symbolic reasoning and self-awareness mechanisms. These approaches represent an important step toward cognitive DTs, but their scope is often limited to specific domains and they do not address large-scale distributed systems.

Liu *et al.* [12] present a comprehensive survey on CDTs and discuss how AI techniques can enhance DT capabilities. In particular, they identify the role of knowledge graphs, reasoning engines, and more recently LLMs in enabling semantic understanding and interaction. Indeed, with the rise of Generative AI, several works have explored the integration of LLMs into DT systems. For instance, Xia *et al.* [29] propose an architecture where LLMs are used to interact with DTs and support reasoning over system data. Similarly, Zhu *et al.* [32] discuss how semantic communication and LLM-based reasoning can enhance DT systems. These works demonstrate the potential of LLMs in enabling explainability, interaction, and cross-domain reasoning.

Recent research has started to explore how AI models, including LLMs, can be distributed across the edge–to-cloud continuum. Surveys such as [10] highlight the potential of deploying lightweight models at the edge to support local processing, while keeping more complex models in the cloud.

Frameworks such as Edge-LLM [3] and hybrid inference approaches [11] propose splitting LLM execution across edge and cloud resources. These approaches enable tasks such as filtering, summarization, and contextual interpretation to be performed locally, reducing communication overhead and improving responsiveness. Similarly, [18] explores optimization strategies for edge–cloud LLM deployment.

These approaches suggest that LLM-based capabilities can be adapted to distributed environments through techniques such as model distillation, specialization, and collaborative inference. In this setting, edge components handle lightweight tasks, while cloud components perform more complex reasoning and decision-making [18]. Nevertheless, even though these works provide important insights into distributed AI, they are not specifically designed for DT systems. In particular, they do not define how distributed AI components should be integrated within a DT architecture, nor how they should interact with local and global twin instances.

The analysis above highlights a clear gap in the current DTs research. Federated

approaches address the decomposition of complex systems into multiple twins, but often lack mechanisms for distributing intelligence and coordinating decisions across them. Cognitive approaches enhance reasoning capabilities, but are typically centralized and do not account for the constraints of distributed CPSs. Research on edge AI and LLM distribution provides enabling technologies, but does not offer a DT-specific architectural framework. As a result, existing solutions tend to address either distribution or cognition, but rarely both in a unified way. This gap motivates the need for a new approach that combines federation and cognition within a single architecture.

3 Federated Cognitive Digital Twin Architecture

The proposed FCDT architecture models the Digital Twin as a federated and cognitive system distributed across the edge-to-cloud continuum. The architecture is based on the distinction between *local twins* and *global twins*, which operate at different levels and provide complementary capabilities. Local twins are associated with specific physical subsystems and are deployed close to them at the edge, while global twins operate in the cloud and reason over the information produced by multiple local twins.

A key principle of the architecture is that part of the DT capabilities is brought closer to the physical system. This is not only a deployment choice, but a functional one. Local twins are not limited to collecting and forwarding data; they provide monitoring, analysis, and a first level of decision-making directly at the subsystem level. This allows them to maintain tighter synchronization with real-world conditions and to react quickly to local changes. The cloud remains essential, but its role shifts from being the central point of intelligence to the level where information are aggregated, correlated, and used for broader reasoning, simulation, and coordination.

Each local twin maintains a continuously updated representation of its subsystem by ingesting data from sensors, devices, and local services. This state forms the basis for real-time analysis, including anomaly detection, short-horizon prediction, and state estimation. Since these activities are performed close to the data source, they can be executed with low latency and without relying on continuous cloud communication, which is crucial in dynamic and time-critical environments.

Based on this analysis, local twins support a first level of decision-making. By combining current state and analytical results, they can take immediate actions within the scope of their subsystem. Although these decisions are inherently local, they provide fast and autonomous responses that would be difficult to achieve in centralized systems. In this way, local twins contribute directly to system responsiveness and resilience.

In addition to these operational capabilities, local twins incorporate lightweight cognitive components, including edge AI and specialized or distilled edge language models. These components support semantic processing tasks such as event extraction, summarization, and contextual interpretation. This enhances the quality of local decisions by enriching numerical data with contextual meaning, which can be useful for human operators or local control logic. At the same time, it enables local twins to transform raw data into structured and compact representations before sharing it. This process, referred to as semantic distillation, reduces communication overhead while preserving relevant information. It can also support privacy-aware data sharing by enabling the transmission of abstracted or anonymized information.

As a result, local twins play a dual role. They act as operational components, responsible for real-time monitoring and local decisions, and as cognitive filters, responsible for transforming raw observations into structured knowledge. These outputs are stored in a shared

data space and constitute the input for higher-level reasoning.

Global twins operate at the cloud level and provide a system-wide perspective. Their role is to reason over the structured information produced by multiple local twins and other global twins, and derive knowledge that cannot be obtained from a single subsystem. Unlike local twins, they do not directly interact with the physical environment, but consume data from the shared data space in the form of events, predictions, and contextualized observations. This abstraction allows them to focus on higher-level and more computationally demanding tasks.

Using these aggregated information, global twins perform system-level analysis, including the identification of correlations across subsystems, detection of global patterns, and long-term prediction. They are also responsible for simulation and the exploration of alternative scenarios. By leveraging the computational resources of the cloud, they can execute complex and time-consuming analyses, such as what-if simulations and the evaluation of coordinated strategies, supporting more informed and anticipatory decision-making.

Global twins are also the main location where the architecture realizes its full cognitive capabilities. They can exploit more powerful AI models, including full-scale LLMs, to support semantic reasoning, explanation, and human interaction. At this level, LLMs are used to interpret aggregated information, relate observations across subsystems, explain system behavior, and assist in understanding the impact of potential actions. In this way, global cognition complements local cognition, providing deeper and broader reasoning over the entire system.

The interaction between local and global twins is mediated by a *shared data space*, which acts as the integration layer of the architecture. This component enables information exchange and coordination without tightly coupling the twins. Instead of direct communication, all components interact through a common information space where structured data, events, summaries, and inferred knowledge are stored and accessed. This design supports decoupling and scalability, as twins do not need to be aware of each other's internal structure.

Through this mechanism, local twins continuously generate structured information and propagate it to the data space. Global twins consume this information, reason over the system state, and produce higher-level insights, strategies, or recommendations. These can then be propagated back to the edge, where local twins use them to refine their behavior. This establishes a bidirectional flow of information, combining bottom-up knowledge propagation with top-down coordination.

Overall, the proposed architecture defines a distributed organization of intelligence. Instead of concentrating all capabilities in the cloud, DT intelligence is explicitly divided across levels according to their role and constraints. At the edge, DT intelligence is lightweight, fast, and closely tied to the physical system, enabling local autonomy and immediate reactions. In the cloud, DT intelligence is more computationally intensive and supports system-level understanding, coordination, and explanation. These capabilities are complementary rather than redundant.

The overall proposed architecture simplifies the engineering of complex DT systems by decomposing the CPS into multiple coordinated twins, improves responsiveness by enabling local decisions, reduces communication through semantic distillation, and supports privacy-aware data sharing. At the same time, it enables advanced reasoning by combining the outputs of multiple twins with simulation and LLM-based interpretation.

4 Conclusion and Future Work

This paper introduced the **Federated Cognitive Digital Twin**, a unified architecture that combines federation and cognition to support DTs in large-scale and distributed CPSs. The proposed approach addresses a key gap in current research. On the one hand, federated DT approaches improve structural decomposition and scalability, but often leave intelligence concentrated in centralized cloud components. On the other hand, cognitive DT approaches enhance semantic reasoning and explainability, but are typically designed as centralized solutions and therefore remain difficult to apply in latency-sensitive and distributed settings.

The FCDT architecture bridges these two lines of work by explicitly distributing DT capabilities across the edge-to-cloud continuum. In the proposed design, local twins operate close to the physical subsystems and provide real-time monitoring, analysis, first-level decision-making, and lightweight cognitive processing. Global twins operate in the cloud and support system-level reasoning, simulation, explanation, and coordination across multiple subsystems. Their interaction through a shared data space enables loose coupling, scalable integration, and bidirectional knowledge exchange. In this way, the architecture combines local autonomy and responsiveness with global semantic reasoning and coordinated decision support.

Overall, the proposed architecture contributes a coherent way to rethink DT engineering in distributed environments. It supports the decomposition of complex CPSs into multiple coordinated twins, reduces the drawbacks of centralized architectures, and extends DTs beyond data processing toward more meaningful and explainable decision support. For these reasons, we believe that FCDT represents a promising architectural direction for DTs in scenarios such as smart cities and other complex distributed systems.

As future work, we plan to refine the architecture along both engineering and evaluation dimensions. First, we will further formalize the responsibilities, interfaces, and interaction patterns of local and global twins, including the role of the shared data space and the coordination mechanisms between edge and cloud components. Second, we will investigate model allocation strategies for deciding which cognitive capabilities should remain at the edge and which should be delegated to the cloud, considering trade-offs among latency, bandwidth, computational cost, privacy, and quality of reasoning. Third, we plan to develop a proof-of-concept implementation and assess the architecture in realistic distributed scenarios, with particular attention to responsiveness, resilience, communication overhead, and the usefulness of semantic distillation and LLM-based reasoning for decision support.

In addition, future work should examine how the proposed architecture can support stronger interoperability and governance. This includes studying semantic integration mechanisms across heterogeneous twins, privacy-aware information sharing, and explainability techniques that make system-level recommendations more transparent to human operators. Finally, we aim to investigate tool support and model-driven engineering techniques for simplifying the design, deployment, and evolution of FCDT-based systems, so that the proposed architecture can be adopted more systematically in practice.

Acknowledgment

This work is supported by the Swedish Agency for Innovation Systems through the project "iSecure: Developing Predictable and Secure IoT for Autonomous Systems" (2023-01899), by the Key Digital Technologies Joint Undertaking through the project "MATISSE: Model-based engineering of digital twins for early verification and validation of industrial systems" (101140216), and by the Clean Energy Transition Partnership through the project “FLEXI:

Human-centered AI and digital twin powered energy system integration for flexibility markets" (101069750).

References

1. Sailesh Abburu, Arne J. Berre, Michael Jacoby, Dumitru Roman, Ljiljana Stojanovic, and Nenad Stojanovic. Cognitwin – hybrid and cognitive digital twins for the process industry. In *2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, pages 1–8, 2020. doi:10.1109/ICE/ITMC49519.2020.9198403.
2. Alessio Bucaioni, Romina Eramo, Luca Berardinelli, Hugo Bruneliere, Benoit Combemale, Djamel Eddine Khelladi, Vittoriano Muttillo, Andrey Sadovykh, and Manuel Wimmer. Multi-partner project: A model-driven engineering framework for federated digital twins of industrial systems (matisse). In *2025 Design, Automation & Test in Europe Conference (DATE)*, pages 1–6. IEEE, 2025.
3. Fenglong Cai, Dong Yuan, Zhe Yang, and Lizhen Cui. Edge-llm: A collaborative framework for large language model serving in edge computing. In *2024 IEEE International Conference on Web Services (ICWS)*, pages 799–809, 2024. doi:10.1109/ICWS62655.2024.00099.
4. Alessandra De Benedictis, Nicola Mazzocca, Alessandra Somma, and Carmine Strigaro. Digital twins in healthcare: An architectural proposal and its application in a social distancing case study. *IEEE Journal of Biomedical and Health Informatics*, 27(10):5143–5154, 2023. doi:10.1109/JBHI.2022.3205506.
5. Kubra Duran, Lal Verda Cakir, Yagmur Yigit, Khayal Huseynov, Sushmitha Ram Kusu, Mehmet Ali Ertürk, and Berk Canberk. Toward digital twin-as-a-service (dtaas) platforms: A survey on architecture, design requirements, and performance metrics. *IEEE Communications Surveys & Tutorials*, 28:1845–1878, 2026. doi:10.1109/COMST.2025.3635582.
6. Enxhi Ferko, Alessio Bucaioni, and Moris Behnam. Supporting technical adaptation and implementation of digital twins in manufacturing. In *International Conference on Information Technology-New Generations*, pages 181–189. Springer, 2012.
7. Enxhi Ferko, Alessio Bucaioni, and Moris Behnam. Architecting digital twins. *IEEE Access*, 10:50335–50350, 2022. doi:10.1109/ACCESS.2022.3172964.
8. Enxhi Ferko, Alessio Bucaioni, Patrizio Pelliccione, and Moris Behnam. Analysing interoperability in digital twin software architectures for manufacturing. In *European conference on software architecture*, pages 170–188. Springer, 2023.
9. Enxhi Ferko, Alessio Bucaioni, Patrizio Pelliccione, and Moris Behnam. Standardisation in digital twin architectures in manufacturing. In *2023 IEEE 20th International Conference on Software Architecture (ICSA)*, pages 70–81. IEEE, 2023.
10. Othmane Friha, Mohamed Amine Ferrag, Burak Kantarci, Burak Cakmak, Arda Ozgun, and Nassira Ghoualmi-Zine. Llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. *IEEE Open Journal of the Communications Society*, 5:5799–5856, 2024. doi:10.1109/OJCOMS.2024.3456549.
11. Zixu Hao, Huiqiang Jiang, Shiqi Jiang, Ju Ren, and Ting Cao. Hybrid slm and llm for edge-cloud collaborative inference. In *Proceedings of the Workshop on Edge and Mobile Foundation Models*, EdgeFM '24, page 36–41, New York, NY, USA, 2024. Association for Computing Machinery. doi:10.1145/3662006.3662067.
12. Yangyang Liu, Tang Ji, Xiangyu Guo, Xun Xu, and Jan Polzer. A survey of cognitive digital twin and the potential use of llms. *Manufacturing Letters*, 44:1242–1253, 2025. 53rd SME North American Manufacturing Research Conference (NAMRC 53). doi:10.1016/j.mfglet.2025.06.144.
13. Andrea Manzi, Raul Bardaji, Ivan Rodero, Germán Moltó, Sandro Fiore, Isabel Campos, Donatello Elia, Francesco Sarandrea, A. Paul Millar, Daniele Spiga, Matteo Bunino, Gabriele Accarino, Lorenzo Asprea, Samuel Bernardo, Miguel Caballer, Charis Chatzikyriakou, Diego Ciangottini, Michele Claus, Andrea Cristofori, Davide Donno, Emanuele Donno, Iacopo

Ferrario, Massimiliano Fronza, Alexander Jacob, Javad Komijani, Marina Krstic Marinkovic, Federica Legger, Ivan Palomo, Estíbaliz Parcero, Rakesh Sarma, Gaurav Sinha Ray, Sara Vallero, and Juraj Zvolensky. intertwin: Advancing scientific digital twins through ai, federated computing and data. *Future Generation Computer Systems*, 179:108312, 2026. doi:10.1016/j.future.2025.108312.

14 Judith Michael, Loek Cleophas, Steffen Zschaler, Tony Clark, Benoit Combemale, Thomas Godfrey, Djamel Eddine Khelladi, Vinay Kulkarni, Daniel Lehner, Bernhard Rumpe, Manuel Wimmer, Andreas Wortmann, Shaukat Ali, Balbir Barn, Ion Barosan, Nelly Bencomo, Francis Bordeleau, Georg Grossmann, Gabor Karsai, Oliver Kopp, Bernhard Mitschang, Paula Muñoz Ariza, Alfonso Pierantonio, Fiona A. C. Polack, Matthias Riebisch, Holger Schlingloff, Markus Stumptner, Antonio Vallecillo, Mark van den Brand, and Hans Vangheluwe. Model-driven engineering for digital twins: Opportunities and challenges. *Systems Engineering*, 28(5):659–670, 2025. doi:10.1002/sys.21815.

15 Roberto Minerva, Gyu Myoung Lee, and Noël Crespi. Digital twin in the iot context: A survey on technical features, scenarios, and architectural models. *Proceedings of the IEEE*, 108(10):1785–1824, 2020. doi:10.1109/JPROC.2020.2998530.

16 Alexios Papacharalampopoulos, Dionysios Christopoulos, Olga Maria Karagianni, and Panagiotis Stavropoulos. Federation in digital twins and knowledge transfer: Modeling limitations and enhancement. *Machines*, 12(10), 2024. doi:10.3390/machines12100701.

17 Renuka Prasad Pasupulati and Jordan Shropshire. Analysis of centralized and decentralized cloud architectures. In *SoutheastCon 2016*, pages 1–7, 2016. doi:10.1109/SECON.2016.7506680.

18 Kunal Rao, Giuseppe Coviello, Priscilla Benedetti, Ciro Giuseppe De Vita, Gennaro Mellone, and Srimat Chakradhar. Eco-llm: Llm-based edge cloud optimization. In *Proceedings of the 2024 Workshop on AI For Systems*, AI4Sys '24, page 7–12. Association for Computing Machinery, 2024. doi:10.1145/3660605.3660941.

19 Daniel Rosendo, Alexandru Costan, Patrick Valduriez, and Gabriel Antoniu. Distributed intelligence on the edge-to-cloud continuum: A systematic literature review. *Journal of Parallel and Distributed Computing*, 166:71–94, 2022. doi:10.1016/j.jpdc.2022.04.004.

20 Mahadev Satyanarayanan. The emergence of edge computing. *Computer*, 50(1):30–39, 2017. doi:10.1109/MC.2017.9.

21 Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, 2016. doi:10.1109/JIOT.2016.2579198.

22 Inderjeet Singh, Eleonore Vissol-Gaudin, Andikan Otung, and Motoyoshi Sekiya. Learning to collaborate: An orchestrated-decentralized framework for peer-to-peer llm federation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(30):25472–25480, Mar. 2026. doi:10.1609/aaai.v40i30.39742.

23 Alessandra Somma, Domenico Amalfitano, Alessio Bucaioni, and Alessandra De Benedictis. A model-driven approach for engineering mobility digital twins: The bologna case study. *Information and Software Technology*, page 107863, 2025.

24 Alessandra Somma, Domenico Amalfitano, Alessandra De Benedictis, and Patrizio Pelliccione. Twinarch: A digital twin reference architecture. *Journal of Systems and Software*, 231:112613, 2026. doi:/10.1016/j.jss.2025.112613.

25 Fei Tao, Bin Xiao, Qinglin Qi, Jiangfeng Cheng, and Ping Ji. Digital twin modeling. *Journal of Manufacturing Systems*, 64:372–389, 2022. doi:10.1016/j.jmsy.2022.06.015.

26 Christian Vergara, Rami Bahsoon, Georgios Theodoropoulos, Wendy Yanez, and Nikos Tziritas. Federated digital twin. In *2023 IEEE/ACM 27th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*, pages 115–116, 2023. doi:10.1109/DS-RT58998.2023.00027.

27 Christian Roberto Vergara, Georgios Theodoropoulos, Rami Bahsoon, Wendy Yanez, and Nikos Tziritas. Federated digital twins as an enabling technology for collaborative decision-

making. In *Proceedings of the 38th ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM-PADS '24, page 67–68. Association for Computing Machinery, 2024. doi:10.1145/3615979.3662152.

28 Christian Vergara-Marcillo, Rami Bahsoon, Nikos Tziritas, and Georgios Theodoropoulos. A connectionist approach to federated digital twins. In *Computational Science – ICCS 2025*, pages 60–74, Cham, 2025. Springer Nature Switzerland.

29 Yuchen Xia, Nasser Jazdi, and Michael Weyrich. An architecture for integrating large language models with digital twins and automation systems. In *2025 IEEE 30th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–8, 2025. doi:10.1109/ETFA65518.2025.11205636.

30 Nan Zhang, Rami Bahsoon, and Georgios Theodoropoulos. Towards engineering cognitive digital twins with self-awareness. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3891–3891, 2020. doi:10.1109/SMC42975.2020.9283357.

31 Xiaochen Zheng, Jinzhi Lu, and Dimitris Kiritsis. The emergence of cognitive digital twin: vision, challenges and opportunities. *International Journal of Production Research*, 60(24):7610–7632, 2022. doi:10.1080/00207543.2021.2014591.

32 Fang Zhu, Jiayuan Chen, Junjie Wen, Yuye Yang, Changyan Yi, Yun Tie, Peng Zhang, Jun Cai, Dusit Niyato, and Mohsen Guizani. From data mirror to smart copilot: A survey on nextg semantic communication for propelling digital twin world into cognitive stage. *IEEE Communications Surveys & Tutorials*, 28:4915–4947, 2026. doi:10.1109/COMST.2026.3665395.