Cohabitation of Intelligence and Systems: Towards Self-reference in Digital Anatomies

Andrea Morichetta[‡], Anna Lackinger[‡], Schahram Dustdar[‡],
[‡]TU Wien, {a.morichetta, a.lackinger, dustdar}@dsg.tuwien.ac.at

2024 IEEE International Conference on Service-Oriented System Engineering (SOSE) | 979-8-3315-3958-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/SOSE62363.2024.00018

***Abstract*—The modern computing scenario of the Computing Continuum exhibits large and complex applications with heterogeneous requirements running on distributed infrastructure. Still, when it comes to coordinating and controlling such applications and infrastructures, it is common to rely on centralized or ad-hoc solutions. While these approaches are robust, scaling management solutions, managing local changes, and having a holistic perspective can be challenging. Additionally, they could be better suited for addressing new problems in dynamic environments. Therefore, new approaches are needed. In this paper, we present DICT, a novel method for managing the Computing Continuum, i.e., the infrastructure and the applications. The proposed approach encompasses a series of modules for automatic management. The core idea is to develop a method for applying the intents coming from the infrastructure and application managers in an autonomic and dynamic way. The modules can communicate through coordinators that take observable inputs and send them back predictions on the next actions to take. These coordinators have the role of summarizing the sensed observation and extracting high-level information in light of the AI advancement that shows how discrete space representation of inputs improves generalization. Thus, they can have models that build their own semantics and "language." We envision that, through DICT, both the application and the infrastructure management will only have to specify high-level intents and not focus on defining encoded and difficult-to-change strategies.**

***Index Terms*—Computing Continuum, Holistic Management, Distributed Intelligence, Intent, Coordination**

I. INTRODUCTION

Enterprise systems have reached the capability of providing pervasive and distributed servers to millions or billions of customers [1]. However, this comes at the price of unprecedented complexity complexity, including a more sophisticated logic. Moreover, the systems operate on infrastructures that span from the IoT to the Cloud, in what is called the "Computing Continuum [2]." The Computing Continuum exhibits heterogeneity in the device types, e.g., IoT devices can be sensors or UAVs [3], and subsets of the Edge or Cloud nodes might feature GPU or AI accelerators. Furthermore, it features intricate interactions. However, solutions for managing this scenario are still in their infancy.

The key challenge is how to guarantee a holistic management [4] of applications and infrastructure, given system-wide and dynamic goals. In particular, infrastructure providers solve the problem by stipulating agreements with the enterprises. However, this approach can be too rigid to adapt to changes in requirements or objectives. Furthermore, the management of the application on and with the infrastructure is centralized. Despite the solid ingenuity of these solutions [5], the monolithic structure of the strategies suffers from the complexity of the Computing Continuum and the dynamicity of the applications. Some approaches aim at breaking up the monolithic orchestration, going towards layered or decentralized approaches [6], [7], [8], [9]. Still, most approaches, despite offering good results in various management tasks [10], [11], [12], are just a set of disconnected models. Therefore, is difficult to have an organic scaling and distribution of the management, both flexible and capable of controlling local changes.

In this work, we present a novel problem setting and framework that enables managing the computing continuum in a holistic and self-automating way. Our key insight, inspired by theories and research from other fields, including cognitive science, biology, physics, and chemistry, is that it is possible to see complex systems with big, varying data from variegated and dynamic components as organisms. In particular, one organism can be seen as composed of a set of specialist modules. Each of these modules acts locally on their portion of inputs. Managing a system composed of independent, specialist modules means having the possibility of generalizing new information, objectives, and applications. This task requires finding communication methods for sharing and improving the modules' knowledge. We propose a novel problem setting where each requirement of an enterprise application or the infrastructure takes the form of a high-level requirement, i.e., intent [13], [14], [15]. Each intent translates into quantifiable measures, e.g., *efficiency*, or *availability*. These measures control the action of low-level strategies. Their interoperation and association form the edges and nodes of a graph. Contextually, each measure relates to some policy or objective for an application and of a subset of the infrastructure, what we call a "*resource group*." These measures need to be adjusted to balance the overall functioning of the system with the global objective of reaching equilibrium.

Within our novel problem setting, we envision DICT (Figure 1), a framework that enables a decentralized and self-automating management of the Computing Continuum. Our blueprint defines the structure of this management. We conceptualize a system composed of multiple independent agents, each specializing in interacting with the underlying infrastructure. In particular, we introduce DICT modules. They manage the agents, taking care of their communication and coordination at different speeds [16], i.e., at various levels of

979-8-3315-3958-0/24/$31.00 ©2024 IEEE

DOI: 10.1109/SOSE62363.2024.00007

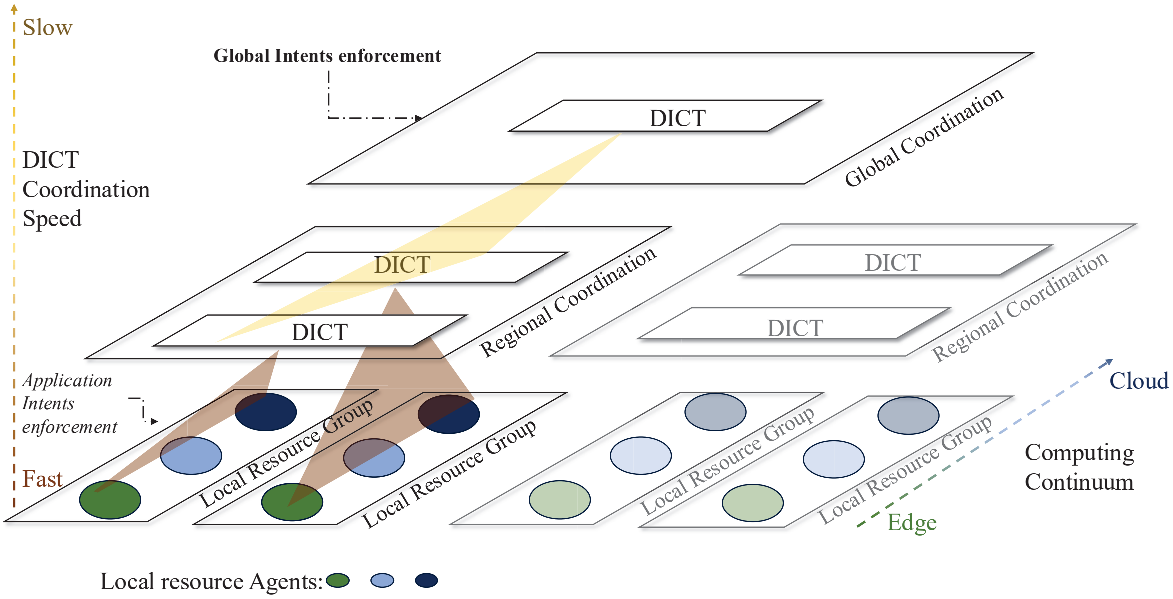

Fig. 1: This figure aims to represent an example of **DICT (Distributed Intelligence Coordination Tool)** management operating at different speeds over the computing continuum infrastructure. It shows how DICT modules are coordinated to enforce application intents from the edge to the cloud. This overview illustrates the varying speeds of coordination, with faster local resource agents (the brown cone) managing immediate tasks at the edge and slower global intents enforcement (yellow cone) occurring in the cloud. This multi-layered approach ensures a seamless and efficient orchestration of resources and tasks across the entire computing continuum, highlighting the dynamic interaction between different system layers.

granularity. We organize DICT modules at different layers of abstraction and connect them according to the intents they manage across various system layers. Each DICT module abstracts the observed input by extracting a higher-level understanding of the system's dynamics and helping predict causally the following states to take the most appropriate actions. We envision that this high-level representation can help deal with the system's uncertainty by offering generalization. This contribution represents one of the main components of the EU Horizon Project, INTEND [17] [1].

The remainder of this paper is structured as follows. Section II provides the background of our study, exploring core concepts from various fields, including cognitive science, biology, physics, and chemistry. In Section III, we discuss advancements in implementing models and tools for complex systems. Section IV introduces DICT, detailing its main components. Finally, Section V aims at providing a viewpoint able to considers the risks associated with AI and machine learning in system automation, emphasizing the importance of safety guarantees and careful implementation.

II. Background

In this Section, we aim at displaying the foundation of our study and characterization. We want to take the core concepts and mechanisms that these approaches are portraying and transfer them to our use case. Specifically, we want to show how the studies of complex systems across fields display similar structure. A complex system model can be seen from the bottom-up perspective as the sum of autonomous agents, from the top-down as a set of mechanisms and message passing patterns that allow to build a global behavior. The topics fundamental to this work that are discussed in this section range from theoretical frameworks that, among other things, take inspiration from the human brain to computational reification, and finally, the overlapping areas of interdisciplinary influences are explored. Thanks to the novel advancements in studies and methods, we believe that having such an approach for the Computing Continuum is realistically achievable.

A. Theoretical Frameworks

a) Global Neuronal Workspace: The Global Workspace Theory is a theory of the brain introduced by Baars [18] and further developed by Dehaene[19], [20]. This theory is interesting because it offers a decentralized model of the brain. The theory envisions that parts of the brain work in autonomy most of the time, performing unconscious tasks. When relevant information is needed, and complex decisions need to be made, a reserved and restricted area of the brain, called the global neuronal workspace, comes into play. This area of the brain collects the inputs that arrive from single specialists, and for the ones that produce relevant information, there is space in the workspace. The information is then aggregated and sent

[1] https://intendproject.eu

back to all the other modules. The model aims to describe the executive control that the brain has over mental processes. As Dehaene describes, it is one of the human attention systems that selects from many possibilities the ones available in the mental operation space.

b) Thinking Fast and Slow: The research on consciousness intertwines with the notorious work on the mind by Daniel Kahneman, "Thinking: Fast and Slow." [21] This work defines, through means of abstraction, a picture of the mind working in two levels, System 1 and System 2. The first performs "unconscious" actions in a "fast" way. More cognitively intense operations require System 2 instead. This system is "slow" as it needs to retrieve the information from a graph of personal knowledge of the world. This work aids us in having a perspective that can consider the speeds of learning and information retrieval and extraction.

c) Active Inference and Multi-Agent Systems: The Principle of minimizing Free Energy (FEP) uncovers the idea of the organism as a set of agents whose role is to minimize uncertainty and optimize their interactions with the environment [22], [23], [24]. Active inference offers mechanisms for acting to minimise free energy. The variational free energy expresses the surprise or uncertainty associated with a belief or uncertainty given a predicted or observed outcome. FEP's idea is that this mathematical structure [25] helps their modeling and self-regulation when applied to large complex systems.

d) Causality: Judea Pearl's framework for Causal Inference provides the foundational framework for modeling cause-and-effect relationships [26]. This work is captivating as it describes how to predict the outcome of some action. Schölkopf and his research group extended this work by identifying the connection between causal inference and machine learning [27]. In particular, this research branch defines how combining the power of causal models with novel Deep Learning methods can lead to better generalization and greater transparency.

B. Computational Reification

a) Recursive Neural Networks: Recursive Neural Networks (RNNs) model the concept of "memory" for neural networks. The idea is to hold relevant information, e.g., anomalies in a time series progression, and share it to enhance knowledge of the current state. This approach is gaining momentum in several works, and it helps represent, with mathematical structure, the idea of having some module that can update the understanding of the system's state.

b) Representation Learning: The last decade of advancements in neural networks sees representation learning as a cornerstone [28]. The core concept is to transform raw input into higher-level, meaningful representations. The extraction of rich information can help downstream tasks to have a better understanding of the problem and generalize over various data. Since the surge of NLP, embeddings have become a key technique for representing discrete variables, such as words in a dictionary, as continuous vectors. Another key method in representation learning is the encoders, which are used for translation and summarization by transforming input data into what is called "latent representations." The idea is to learn to encode a different representation of the data by extracting the relevant information that the input carries, which gives the decoder the capability to understand and act better. Autoencoders guarantee representation and reconstruction. The autoencoder is used to learn the internal structure of some input and encode it in a hidden internal representation. A version of them is the variational autoencoders that use probability. They are generative models that, thanks to the probability capabilities, can better understand the underlying causal relation. This causal understanding leads to better generalization.

c) Cellular Automata and Neural Cellular Automata: Cellular automata is an architecture and a rule for formalizing and representing the self-reproduction in machines. Von Neumann first introduced it, and later, it was reworked by John Conway and Stephen Wolfram [29], [30]. What cellular automata provide is parallelism and the capability to form complex patterns. This approach guarantees the achievement of computation in a non-traditional way.

C. Interdisciplinary Influences

a) Complex Systems Theories: Talking about complex systems science in a paragraph would be oversimplifying. Here, we aim to highlight the perspective of a set of models and principles that can become essential when dealing with the problem of continuum management. Specifically, concepts like self-organization and emergence, both inspired and applied to natural phenomena, have helped formalize the behavior of systems that exhibit complexity, intended as the cases where the whole system behavior is more than the sum of the individual actions of the agents that compose the system. In particular, complex systems research has emphasized the quest for interdisciplinary findings. Certainly, the adoption of complex systems concepts in distributed systems is not new; over the years, contributions from scholars like Bernardo Huberman and Eric Bonabeau have been pivotal. Still, the modern techniques and mathematical frameworks that both fields of cognitive science and artificial intelligence are providing can lead to the application of such paradigms on large-scale systems. Indeed, novel approaches like self-organized criticality, based on concepts like dynamical systems theory, are being applied with success to various fields, from the brain [24] to natural phenomena [31].

b) Surfing Uncertainty: The work of Andy Clark, "Surfing Uncertainty" [32], offers a general and abstract characterization of the nervous system. In particular, Clark takes a top-down perspective. In particular, he presents a hierarchical, action-oriented predictive processing organization. The rationale is that the cognitive system can be represented as a set of levels organized hierarchically. The higher levels have the role of predicting the stimuli coming from the lower levels. The mismatches between the predictions and the stimuli are sent back to the higher levels to improve future inferences. This approach implies the realization of a generative model, specifically the result of a process of Bayesian inference.

The agents that compose this characterization are responsible for generating predictions about the world. In particular, they perform precision-weighting, i.e., extracting the information of the sensory stimuli that are more relevant to make an accurate prediction. In this way, the system builds a model of the world that will eventually grow in accuracy over time.

III. Advancements and Opportunities

Incorporating advances and opportunities in different areas is crucial to creating a versatile and practical solution when developing an intelligent coordination tool. The topics discussed in this section include collective agents' cooperation, message passing among agents, system learning over time, and the design of accountable and responsible frameworks. Complex problems can be solved more effectively by fostering cooperation among multiple agents. Therefore, we first want to introduce solutions for coordinating multiple agents. Since effective communication is key to the success of any distributed system, we then introduce some ideas and solutions that have been proposed in this area. In the next part, we discuss system learning over time, which is essential to ensure that the system performance does not deteriorate. The last part discussed the critical topic of designing accountable and responsible frameworks.

A. Collective Multi-agent Systems implementation

The advent of distributed computing, characterized by a wide continuum spanning from edge devices to cloud infrastructure, demands innovative management strategies to optimize resource allocation and ensure seamless integration. Multi-agent Systems (MAS) demonstrated over time to enable the integration of various goal and logic in an autonomous, yet cooperative way [33]. In the context of virtualized infrastructures, as the Cloud[34] and the Edge [35], MAS proved the maturity to handle complex decision-making processes, especially for task placement and resource allocation. Yet, considering Edge and Cloud in isolation is not enough; therefore, it is essential to develop MASs in the whole Computing Continuum, as a central pillar for evolving to more adaptive, resilient, and scalable architectures. Contingently, providing an efficient and accurate MAS requires leveraging self-regulating algorithms [36] and developing accurate coordination and communication strategies [37].

1) Control-based approaches: When dealing with complex algorithms, it is necessary to coordinate multiple agents and their actions [38]. Both can broadly vary in scope and characteristics (think about the variety of data and functions in a composite scenario: computing, monitoring, managing, etc.). In particular, Reinforcement-learning-based techniques, also called MARL (Multi Agent Reinforcement Learning) [39], offer the most prominent structure for acting on the environment and iteratively improve the management. In particular, new approaches leverage Representation Learning [40] methods for improving the understanding of the observed environment.

a) Swarm intelligence: Swarm intelligence (SI) explores biological principles such as stigmergy and self-organization and highlights the importance of interactions and the effects of parameter changes on collective behavior[41]. In a swarm, the individual organisms, relying on limited and imprecise environmental information, collectively navigate through uncertainties and solve complex problems. This area of research has its origins in the study of the self-organized behavior of social insects and has implications in various fields such as telecommunications and autonomous robotics[41]. Swarm intelligence can be split into insect-based and animal-based algorithms [42], However, this intelligence can also be observed in other environments, e.g. in colonies of bacteria or amoebae, crowds of human beings, and many others [41]. One popular SI algorithm is called Ant Colony Optimization (ACO), which is inspired by the foraging behavior of certain ant species that create pheromone trails to signal optimal paths to their colony mates. This method uses a similar strategy to solve optimization problems in a decentralized environment. Although these approaches can handle complex decision-making processes, managing the scale that the Computing Continuum embeds is still challenging. In addition, such approaches require further development to manage the dynamic nature of the computing continuum and generalize to new requirements from the infrastructure and application managers.

2) Deep Learning-based appraoches: Other approaches explore the use of Representation Learning for controlling MAS, e.g., through shared memory mechanisms [43]. Another possible application of these principles can be seen in the "How2comm" framework [44]. The paper proposes an approach to collaborative perception and aims at optimizing the trade-off between perception accuracy and communication efficiency. Collaborative perception itself aims at integrating sensory data to achieve a more accurate and comprehensive understanding of the environment and more accurate build and reasoning on complex systems. In this context, a strategy focusing on meaning can improve communication redundancy, transmission delay, and collaboration [45]. In this set of techniques graphs [46], [47] play an essential role in highlighting the relationships between agents and the most relevant communication information.

B. Semantic Communication

Communication is not trivial since it requires high-level semantics and regulatory mechanisms. Several works attempt to achieve it, proposing various encoding mechanisms [48]. Semantic communication [49], [50], a pillar for complex systems [51], is focusing on the meaning and context of exchanged information [50], [52], emerges as a pivotal solution for enhancing interoperability and efficiency across the computing continuum [53]. In particular, semantic communication has the potential to address key challenges in real-time, complex distributed systems as the ones that involve autonomous driving.

An open research question is how a designed communication scheme can combine these approaches and guarantee that specialized algorithms can communicate with each other. Indeed, when developing distributed solutions, having robust encoding mechanisms that can handle diverse input becomes essential.

C. System learning over time

a) Learning methods inspired by cognitive psychology, physics, and neuroscience: Several learning approaches have been inspired by how humans learn, to improve the learning ability of artificial intelligence systems. De Melo et al. [54] describe three different approaches, including multimodal learning, continual learning, and embodied learning. Multimodal learning uses redundancy and self-monitoring of multiple sensory inputs such as vision, hearing, and touch to improve learning in humans and deep learning systems. In general, the goal is to integrate different sensor data [55] for more robust task performance and comprehensive training. Continual learning is characterized by the ability to continuously adapt and learn in a changing environment. It inspires deep learning (DL) models to support the learning of an infinite set of tasks using mechanisms such as weight protection, memory systems for repetition, and network modularity to prevent forgetting and enable lifelong learning. Some of these approaches can be categorized under the umbrella of multi-task learning [56]. In this case, the capabilities of generalization to various domains [57] or, better, to shifts in inputs, is essential. Embodied learning, which emphasizes the importance of interactive engagement with the environment for knowledge acquisition, shifts the paradigm towards the development of DL systems that learn from exploratory, multimodal feedback, supported by simulated environments and techniques such as inverse rendering and latent spatial entanglement to understand the properties of the 3D world. In particular, causal learning [58], [59] techniques are offering the discovery of high- level causal variables from low-level observations. These approaches aim at providing better generalization capabilities for deep learning and faster convergence for RL models. They are also essential for continual and multi-task learning, managing to decompose knowledge about the world into indepen- dent and recomposable pieces.

b) Importance of Reasoning in Distributed AI: The challenge is making complex applications collaborate at different hierarchy levels and at different speed [16] and take intelligent decisions. The vast amount of data required to run current algorithms is not sustainable anymore, and systems need to adapt to out-of-distribution data. Therefore, complex environments demand calls for the development of techniques that can combine top-down signals (high-to-low level - cloud-to-edge) with bottom-up ones (from sensors up), consolidating the cloud-edge continuum as a unique ”organism.” A set of recent studies in the deep learning community developed the combination of top-down and bottom-up signals on a more abstract level, inspired by what allegedly happens in our brain [60]. In particular, one of the deep learning techniques proposed in [61] derives from the idea of a constrained shared resource pool in the brain where decisions compete to the access, letting the best configurations of modules win. This bottleneck seems beneficial for improving coordination and communication between specialist models selecting the best set of actions, and various research groups are researching in the same direction [62]. These approaches take part in a wider research quest to develop models capable of offering generalization [63] capabilities under uncertainty. In particular, recent efforts have been done to define the essential step for *modular*, adaptive models [64], based on deep learning techniques. These advancement are essential to have a better understanding on how to build automated, AI tools with more capabilities [51], [65]

The research for this topic involves studying solutions for distributed applications by evaluating them in various scenarios.

IV. System Blueprint

Our aim is to develop a novel federated decision coordinator, the *Distributed Intelligence Coordination Tool* (**DICT**), designed for data pipelines in the computing continuum. The main idea is to enable distributed decision-making that includes local and regional coordinators. Within the framework of the INTEND Project [17], we envision DICT operating within an infrastructure that has global outreach, with a European core. The system itself is dynamic and continually evolving, necessitating distributed control mechanisms. The goal is to move towards self-organizing, self-regulating systems without human intervention.

DICT role is to coordinate intents, i.e., high-level requirements that are defined for data pipeline entities. The system processes these intents by converting them into logical constraints for AI models. Actions are then analyzed by the Autonomic Manager using information from sensors and predefined rules. Conflicts are resolved through DRL, information sharing, and optimized resource allocation.

A. DICT *Coordinator*

We envision the coordination of the computing continuum in a hierarchical, intent-based way. The hierarchical structure, depicted in Figure 2, guarantees that the management and the conflicts are handled at various levels of detail. The *global coordinator* focuses on global intents and policies, making strategic decisions that impact the whole system. It focuses on guaranteeing long-term goals and enforcing a coherent approach across the entire system. Its tasks include balancing and finding trade-offs among high-level intents. The global coordinator manages multiple regional coordinators. *Regional coordinators* manage mid-level, regional decisions. A region is a cluster of logically connected infrastructure or network regions. Each regional coordinator prioritizes regional intent demands while aligning with the directives coming from the top-level/global coordinator. The *local coordinators* manage a restricted portion of the infrastructure. They directly synchronize with the local agents at the machine, data, and network level. The local coordinator guarantees real-time decision-making, responding quickly to local conditions and observ-

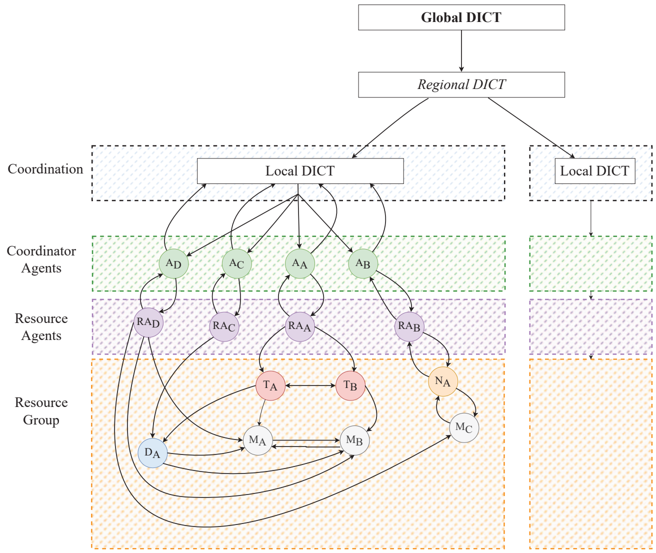

Fig. 2: A simplified representation of the various entities and the hierarchy that constitute the DICT environment and how they interact with each other.

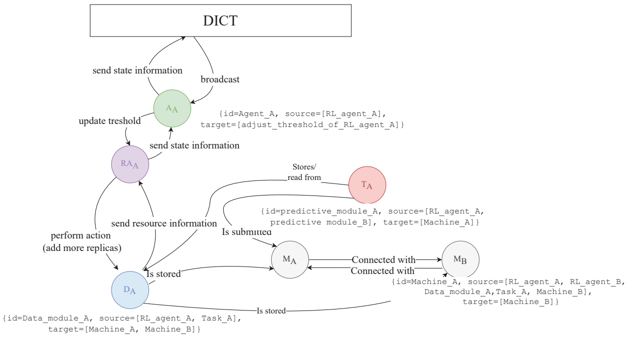
DICT
send state information
broadcast
A_A
{id=Agent_A, source=[RL_agent_A],
target=[adjust_threshold_of_RL_agent_A]}
update treshold
send state information
RA_A
Stores/
read from
T_A
{id=predictive_module_A, source=[RL_agent_A,
predictive module_B], target=[Machine_A]}
send resource information
Is submitted
perform action
(add more replicas)
M_A
Connected with
Connected with
M_B
Is stored
{id=Machine_A, source=[RL_agent_A, RL_agent_B,
Data_module_A,Task_A, Machine_B],
target=[Machine_B]}
D_A
Is stored
{id=Data_module_A, source=[RL_agent_A, Task_A],
target=[Machine_A, Machine_B]}

Fig. 3: Illustrative pipeline for the local DICT coordinator and its interaction with the submodules

able inputs. Its adjustments are targeting the optimization of local performance. At the same time, its decisions are driven by the regional coordinator's rules. Figure 3 depicts a more detailed example of a local coordinator. The local coordinator receives the quantitative award functions from the upper layers and manages one research group with one or more resource dimensions. As depicted, the local coordinator forwards the functions to coordinator agents, and it controls and adjusts award functions based on feedback coming from the coordinator agents.

The operations of the coordinators are based on concepts developed in neuroscience [19], [66] and cognitive studies [20], recently extended in cutting-edge Artificial Intelligence research [61], [63]. The rationale is to consider the management of such complex systems as modular and distributed. The term "modularity" emphasizes the designing of specialized expert systems tailored for specific tasks. This aspect is essential as we do not want to have a unique, monolithic model that is difficult to manage [67], interpret, and maintain. On the contrary, we want specialized modules that are good at predicting, classifying, or making local decisions. "Distribution" enables these modules to communicate and collaborate seamlessly. This aspect entails the selection of a subset of all the available modules to solve specific tasks related to the current intents. Therefore, it enables the concept of composability and reuse of already present solutions, adapting them to the specific needs. Thus, the coordinator guarantees communication and collaboration between the expert modules. As stated in the recent work by Pfeiffer et al. [64], two of the main challenges with multimodal systems are related to developing routing and aggregation functions. The routing function development answers the question "How are active modules selected?" i.e., *which specialized modules it selects for a specific task?*

B. Resource Groups and Agents

For DICT, each coordinator manages a resource group or a series of resource groups at various hierarchy levels. The three main resource types in the INTEND continuum are the *Infrastructure*, *Data*, and *Network*. There are two types of agents, including resource and coordinator agents, which are depicted in Figure 3 and either manage a resource group or a resource agent. *Coordinator agents* receive a quantitative award function from the DICT and forward the function to *resource agents* in the form of controllable constraints or quantitative award functions. Additionally, they control and adjust award functions based on feedback coming from the resource agents. Resource agents perform actions on the resource group they manage and send information to the coordinator agents. The agents are built with principles of modularity, reusability, and composition. Therefore, over time, a resource group agent that, e.g., focuses on autoscaling for a specific application can perform the same task on a new application. This information should be updated in the Knowledge Graph, which is described in the subsection IV-C. At the same time, thanks to the modularity feature, we can introduce a new agent that performs a new task in the system. For example, for a resource group, we add an agent that predicts a machine's response time. This new node will be connected to the target it acts on and to the resource group's coordinator.

C. The Role of the Graph

For DICT, each coordinator manages a resource group or a series of resource groups at various hierarchy levels. Information stored in the Knowledge Graph (KG) includes the data representing the dynamic state and capabilities of agents that the coordinator manages, together with the information regarding the components of the three main classes in the INTEND continuum, which are the *Infrastructure*, *Data*, and *Network*. If a resource group depends on an intent and the coordinator manages it, that should be read from the KG. This information includes the available agents, their type, current states, capacities, and roles within the resource group environment. The DICT tool would also benefit from information about the infrastructure, the network, the application, and the data space, which directly impact the agents' activities. So, when reading from the KG, the edges correspond to each agent's inputs and outputs. When an agent that, e.g., focuses on autoscaling for a specific application performs a task on the application, this information should be updated in the KG (target). At the same time, thanks to the modularity feature, we can introduce a new agent that performs a new task in the system. For example, for a resource group, we add an agent that predicts a machine's response time. This new node will be connected to the target it acts on and to the resource group's coordinator.

V. Navigating AI Risks in System Automation

When we talk of ML or AI, we must consider this branch of science as a reflection of our society [68], [69]. In particular, the fragmentation and automation of procedures in science and society are deeply reflected in ML models. Therefore, when developing such models, we always have to think of the reasons and of which procedures we are trying to automate. Furthermore, we need to be aware that the observation of sensed data does not happen from an acritical and objective perspective, but already the viewpoint the programs we write to collect information are necessarily embodying human conceptualizations [70]. As a consequence, a degree of uncertainty on what might be "uncovered" is inevitable. From the standpoint of the designers of such models for managing complexity and taking the perspective of statistical "language," [71] we need to be aware that finding a straightforward way to encode uncertainty is not trivial and eventually not fully possible. In this scenario, it is also essential to discuss its safety guarantees, especially in deep learning and autonomous systems. Recent contributions [72] are opening questions about how to manage long-term autonomous systems where the current AI models lack epistemic humility, interpretability, and explainability. With concerns about AI drifting or being used maliciously, addressing how to ensure AI remains aligned with human values and controlled is essential. The emphasis could be given on the importance of reward-based training within strict boundaries. Integrating Bayesian inference and causal networks can help create these boundaries and mitigate the risks. In addition, there is the necessity of forming international alliances to address AI safety and work on making recent policies [73], [74], [75] concrete.

VI. Acknowledgment

This work is funded by the HORIZON Research and Innovation Action 101135576 INTEND "Intent-based data operation in the computing continuum." Special thanks to our colleague Alireza Furutanpey for the exchange of views, the discussions, and the feedbacks.

References

[1] "Similarweb: Digital market intelligence platform," https://pro.similarweb.com/#/digitalsuite/webmarketanalysis/home, accessed: 2024-06-13.

[2] S. Dustdar, V. C. Pujol, and P. K. Donta, "On distributed computing continuum systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 4092–4105, 2022.

[3] L. S. Vailshery, "Iot and non-iot connections worldwide 2010–2025," *Stat. March*, 2021.

[4] A. Morichetta, V. C. Pujol, and S. Dustdar, "A roadmap on learning and reasoning for distributed computing continuum ecosystems," in *IEEE EDGE*, 2021.

[5] S. Tuli, S. S. Gill, M. Xu, P. Garraghan, R. Bahsoon, S. Dustdar, R. Sakellariou, O. Rana, R. Buyya, G. Casale *et al.*, "Hunter: Ai based holistic resource management for sustainable cloud computing," *Journal of Systems and Software*, vol. 184, p. 111124, 2022.

[6] M. Schwarzkopf, "The evolution of cluster scheduler architectures," *Retrieved March*, vol. 13, p. 2019, 2016.

[7] M. Abouelyazid and C. Xiang, "Architectures for ai integration in next-generation cloud infrastructure, development, security, and management," *International Journal of Information and Cybersecurity*, vol. 3, no. 1, pp. 1–19, 2019.

[8] T. Pusztai, S. Nastic, A. Morichetta, V. C. Pujol, P. Raith, S. Dustdar, D. Vij, Y. Xiong, and Z. Zhang, "Polaris scheduler: Slo-and topology-aware microservices scheduling at the edge," in *2022 IEEE/ACM 15th International Conference on Utility and Cloud Computing (UCC)*. IEEE, 2022, pp. 61–70.

[9] T. Pusztai, S. Nastic, P. Raith, S. Dustdar, D. Vij, and Y. Xiong, "Vela: A 3-phase distributed scheduler for the edge-cloud continuum," in *2023 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 2023, pp. 161–172.

[10] S. Ilager, R. Muralidhar, and R. Buyya, "Artificial intelligence (ai)-centric management of resources in modern distributed computing systems," in *2020 IEEE Cloud Summit*. IEEE, 2020, pp. 1–10.

[11] A. Morichetta, T. Pusztai, D. Vij, V. C. Pujol, P. Raith, Y. Xiong, S. Nastic, S. Dustdar, and Z. Zhang, "Demystifying deep learning in predictive monitoring for cloud-native slos," in *2023 IEEE 16th International Conference on Cloud Computing (CLOUD)*. IEEE, 2023, pp. 1–11.

[12] A. Morichetta, V. C. Pujol, S. Nastic, S. Dustdar, D. Vij, Y. Xiong, and Z. Zhang, "Polarisprofiler: A novel metadata-based profiling approach for optimizing resource management in the edge-cloud continnum," in *2023 18th Annual System of Systems Engineering Conference (SOSE)*, 2023.

[13] N. Filinis, I. Tzanettis, D. Spatharakis, E. Fotopoulou, I. Dimolitsas, A. Zafeiropoulos, C. Vassilakis, and S. Papavassiliou, "Intent-driven orchestration of serverless applications in the computing continuum," *Future Generation Computer Systems*, vol. 154, pp. 72–86, 2024.

[14] A. Morichetta, N. Spring, P. Raith, and S. Dustdar, "Intent-based management for the distributed computing continuum," in *2023 IEEE International Conference on Service-Oriented System Engineering (SOSE)*. IEEE, 2023, pp. 239–249.

[15] J. Spillner, J. F. Borin, and L. F. Bittencourt, "Intent-based placement of microservices in computing continuums," in *Future Intent-Based Networking: On the QoS Robust and Energy Efficient Heterogeneous Software Defined Networks*. Springer, 2021, pp. 38–50.

[16] J. Medrano, K. Friston, and P. Zeidman, "Linking fast and slow: the case for generative models," *Network Neuroscience*, vol. 8, no. 1, pp. 24–43, 2024.

[17] D. Firmani, F. Leotta, J. G. Mathew, J. Rossi, L. Balzotti, H. Song, D. Roman, R. Dautov, E. J. Husom, S. Sen, V. Balionyte-Merle, A. Morichetta, S. Dustdar, T. Metsch, V. Frascolla, A. Khalid, G. Landi, J. Brenes, I. Toma, R. Szabó, C. Schaefer, C. Udroiu, A. Ulisses, V. Pietsch, S. Akselsen, A. Munch-Ellingsen, I. Pavlova, H.-G. Kim, C. Kim, B. Allen, S. Kim, and E. Paulson, "Intend: Intent-based data operation in the computing continuum," in *Proceedings of the 36th International Conference on Advanced Information Systems Engineering (CAiSE 2024), Limassol, Cyprus, June 10-14, 2024*. Springer, 2024.

[18] B. J. Baars, "Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience," *Progress in Brain Research*, vol. 150, pp. 45–53, 2005.

[19] S. Dehaene, H. C. Lau, and S. Kouider, "What is consciousness, and could machines have it?" *Science*, vol. 358, pp. 486 – 492, 2017.

[20] S. Dehaene, *How We Learn: The New Science of Education and the Brain*. Penguin UK, 2020.

[21] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[22] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo, "Active inference: a process theory," *Neural computation*, vol. 29, no. 1, pp. 1–49, 2017.

[23] T. Parr, G. Pezzulo, and K. J. Friston, *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

[24] J. S. Bettinger and K. J. Friston, "Conceptual foundations of physiological regulation incorporating the free energy principle and self-organized criticality," *Neuroscience & Biobehavioral Reviews*, p. 105459, 2023.

[25] M. Andrews, "The math is not the territory: navigating the free energy principle," *Biology & Philosophy*, vol. 36, no. 3, p. 30, 2021.

[26] J. Pearl, "Causal inference in statistics: An overview," 2009.

[27] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[28] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[29] S. Wolfram, *Cellular automata and complexity: collected papers*. crc Press, 2018.

[30] M. Ghosh, R. Kumar, M. Saha, and B. K. Sikdar, "Cellular automata and its applications," in *2018 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. IEEE, 2018, pp. 52–56.

[31] H. Hoffmann and D. W. Payton, "Optimization by self-organized criticality," *Scientific reports*, vol. 8, no. 1, p. 2358, 2018.

[32] A. Clark, *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press, 2015.

[33] R. Kishore, H. Zhang, and R. Ramesh, "Enterprise integration using the agent paradigm: foundations of multi-agent-based integrative business information systems," *Decision support systems*, vol. 42, no. 1, pp. 48–78, 2006.

[34] X. Gao, R. Liu, and A. Kaushik, "Hierarchical multi-agent optimization for resource allocation in cloud computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 3, pp. 692–707, 2020.

[35] Y. Li, X. Zhang, T. Zeng, J. Duan, C. Wu, D. Wu, and X. Chen, "Task placement and resource allocation for edge machine learning: a gnn-based multi-agent reinforcement learning paradigm," *IEEE Transactions on Parallel and Distributed Systems*, 2023.

[36] S. S. Gill, M. Xu, C. Ottaviani, P. Patros, R. Bahsoon, A. Shaghaghi, M. Golec, V. Stankovski, H. Wu, A. Abraham *et al.*, "Ai for next generation computing: Emerging trends and future directions," *Internet of Things*, vol. 19, p. 100514, 2022.

[37] D. Simões, N. Lau, and L. P. Reis, "Multi-agent actor centralized-critic with communication," *Neurocomputing*, vol. 390, pp. 40–56, 2020.

[38] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003.

[39] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[40] X. Wang, Z. Zhang, and W. Zhang, "Model-based multi-agent reinforcement learning: Recent progress and prospects," *arXiv preprint arXiv:2203.10603*, 2022.

[41] S. Garnier, J. Gautrais, and G. Theraulaz, "The biological principles of swarm intelligence," *Swarm intelligence*, vol. 1, pp. 3–31, 2007.

[42] A. Chakraborty and A. K. Kar, "Swarm intelligence: A review of algorithms," *Nature-inspired computing and optimization: Theory and applications*, pp. 475–494, 2017.

[43] S. Wang, B. Hindman, and I. Stoica, "In reference to rpc: It's time to add distributed memory," in *Proceedings of the Workshop on Hot Topics in Operating Systems*, 2021, pp. 191–198.

[44] D. Yang, K. Yang, Y. Wang, J. Liu, Z. Xu, R. Yin, P. Zhai, and L. Zhang, "How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[45] Y. Hu, X. Pang, X. Qin, Y. C. Eldar, S. Chen, P. Zhang, and W. Zhang, "Pragmatic communication in multi-agent collaborative perception," *arXiv preprint arXiv:2401.12694*, 2024.

[46] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 4106–4115.

[47] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.

[48] H. Seo, J. Park, M. Bennis, and M. Debbah, "Semantics-native communication with contextual reasoning," *arXiv preprint arXiv:2108.05681*, 2021.

[49] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," *arXiv preprint arXiv:2201.01389*, 2021.

[50] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, "Towards a theory of semantic communication," in *2011 IEEE Network Science Workshop*. IEEE, 2011, pp. 110–117.

[51] P. Agrawal, C. Tan, and H. Rathore, "Advancing perception in artificial intelligence through principles of cognitive science," *arXiv preprint arXiv:2310.08803*, 2023.

[52] E. Uysal, O. Kaya, A. Ephremides, J. Gross, M. Codreanu, P. Popovski, M. Assaad, G. Liva, A. Munari, B. Soret *et al.*, "Semantic communications in networked systems: A data significance perspective," *IEEE Network*, vol. 36, no. 4, pp. 233–240, 2022.

[53] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2022.

[54] C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins, "Next-generation deep learning based on simulators and synthetic data," *Trends in cognitive sciences*, 2022.

[55] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, "Scientific machine learning through physics–informed neural networks: Where we are and what's next," *Journal of Scientific Computing*, vol. 92, no. 3, p. 88, 2022.

[56] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[57] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2022.

[58] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[59] J. Berrevoets, K. Kacprzyk, Z. Qian, and M. van der Schaar, "Causal deep learning," *arXiv preprint arXiv:2303.02186*, 2023.

[60] S. Mittal, A. Lamb, A. Goyal, V. Voleti, M. Shanahan, G. Lajoie, M. Mozer, and Y. Bengio, "Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6972–6986.

[61] A. Goyal, A. Didolkar, A. Lamb, K. Badola, N. R. Ke, N. Rahaman, J. Binas, C. Blundell, M. Mozer, and Y. Bengio, "Coordination among neural modules through a shared global workspace," *arXiv preprint arXiv:2103.01197*, 2021.

[62] R. VanRullen and R. Kanai, "Deep learning and the global workspace theory," *Trends in Neurosciences*, vol. 44, no. 9, pp. 692–704, 2021.

[63] A. Goyal and Y. Bengio, "Inductive biases for deep learning of higher-level cognition," *Proceedings of the Royal Society A*, vol. 478, no. 2266, p. 20210068, 2022.

[64] J. Pfeiffer, S. Ruder, I. Vulić, and E. M. Ponti, "Modular deep learning," *arXiv preprint arXiv:2302.11529*, 2023.

[65] P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji *et al.*, "Consciousness in artificial intelligence: insights from the science of consciousness," *arXiv preprint arXiv:2308.08708*, 2023.

[66] C. J. Whyte and R. Smith, "The predictive global neuronal workspace: A formal active inference model of visual consciousness," *Progress in neurobiology*, vol. 199, p. 101918, 2021.

[67] C.-J. M. Liang, Z. Fang, Y. Xie, F. Yang, Z. L. Li, L. L. Zhang, M. Yang, and L. Zhou, "On modular learning of distributed systems for predicting {End-to-End} latency," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, 2023, pp. 1081–1095.

[68] M. Andrews, "The immortal science of ml: Machine learning & the theory-free ideal," *Preprint at https://rgdoi. net/10.13140/RG*, vol. 2, no. 28311.75685, 2023.

[69] ——, "The devil in the data: Machine learning & the theory-free ideal," 2023.

[70] A. Birhane, "Automating ambiguity: Challenges and pitfalls of artificial intelligence," *arXiv preprint arXiv:2206.04179*, 2022.

[71] C. Gruber, P. O. Schenk, M. Schierholz, F. Kreuter, and G. Kauermann, "Sources of uncertainty in machine learning–a statisticians' view," *arXiv preprint arXiv:2305.16703*, 2023.

[72] M. K. Cohen, N. Kolt, Y. Bengio, G. K. Hadfield, and S. Russell, "Regulating advanced artificial agents," *Science*, vol. 384, no. 6691, pp. 36–38, 2024.

[73] H.-L. E. G. on Artificial Intelligence, “Ethics guidelines for trustworthy ai,” 2019. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

[74] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition*. IEEE, 2019. [Online]. Available: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

[75] T. Madiega, “Artificial intelligence act,” *European Parliament: European Parliamentary Research Service*, 2021.