A roadmap on learning and reasoning for distributed computing continuum ecosystems

Andrea Morichetta
Distributed Systems Group, TU Wien
Vienna, Austria
morichetta@dsg.tuwien.ac.at

Victor Casamayor Pujol
Distributed Systems Group, TU Wien
Vienna, Austria
v.casamayor@dsg.tuwien.ac.at

Schahram Dustdar
Distributed Systems Group, TU Wien
Vienna, Austria
dustdar@dsg.tuwien.ac.at

***Abstract*—A captivating set of hypotheses from the field of neuroscience suggests that human and animal brain mechanisms result from few powerful principles. If proved to be accurate, these assumptions could open a deep understanding of the way humans and animals manage to cope with the unpredictability of events and imagination. Modern distributed systems also deal with uncertain scenarios, where environments, infrastructures, and applications are widely diverse. In the scope of Edge-Fog-Cloud computing, leveraging these neuroscience-inspired principles and mechanisms could aid in building more flexible solutions able to generalize over different environments. In this work, we focus on the approaches that center on high-level, general strategies, like the Free Energy Principle and Global Neuronal Workspace theories. The goal of exploring these techniques is to introduce principles that can potentially help us build distributed systems able to jointly work on the whole computing continuum, from the Edge to the Cloud, with self-adapting capabilities, i.e., dealing with uncertainty and the need for generalization, which is currently an open issue.**

I. Introduction

The development of software systems that perform on multiple computing tiers, including IoT, Edge, Fog, and Cloud, promises new opportunities for applications that require features provided by a specific tier. As we envision increasingly complex applications, deriving features only from a specific computing layer is insufficient, demanding enlarging the perspective on all the computing tiers [1]. This scenario entails a new paradigm, i.e., the aggregation of all computing tiers, also known as the computing continuum.

One of the first issues that arise when dealing with applications requiring a computing continuum is how to manage them. The concurrent execution of an application on the entire computing continuum and its dependency on the underlying infrastructure makes it virtually impossible to specify its management solely on the application software. The methodologies developed for the Cloud tier, such as elasticity, do not adequately fit on the other tiers. Therefore, we aim at proposing a set of novel methodologies to manage distributed systems of the computing continuum.

To engage in this new endeavor, we need to highlight the complexity of these systems inherent in their distributed fashion. This scenario invalidates the idea of having single-tier, centralized management, such as the cloud. On the contrary, it pushes towards distributing it along the tiers, addressing management from a holistic perspective [2]. This view is a complete game-changer for distributed computing systems and calls for exploring scientific methods and technologies that currently deal with systems that share these key characteristics, again distributed and complex, such as ecosystems.

If we look beyond computer science, several approaches in neuroscience deal with the brain and human body behaviors, modeling them as complex and distributed systems [3], [4], [5], [6]. If we abstract some human biology concepts and cognitive reasoning, we can form a functional and communicative perspective, describing them as distributed systems [7]. Friston [3] proposes a cognitive neuroscience theory, modeling the adaptive behavior of the brain under what he calls the *Free Energy Principle (FEP)*. It theorizes that human cognition tries to minimize the difference between predicting the environment and its actual observation. Friston describes this process as minimizing the *Free Energy* - in other words, the amount of *surprise* humans feel when the perceived signals of the environment do not match their prediction.

This article aims at envisioning a blueprint for adaptive and self-organizing distributed systems of the computing continuum. To do so, we draw the requirements for a system inspired by FEP-related concepts. Then, we present a possible set of technologies for implementing such a system. Figure 1 shows our research roadmap, depicting the main requirement for using the FEP. Proposing *generative models* for interoception, proprioception, and exteroception, that is, generative models for the *internal behavior* of the system, for the *acting capacity* of the system, and the system's *environmental behavior*, respectively. To achieve these three generative models, three key scientific methods and technologies, apart from the knowledge on distributed systems, are fundamental: causal inference, deep learning, and semantic communication.

Causal inference [8], i.e., extracting a causal connection given conditions linked to an event, plays a fundamental role in generalizing the dynamics of the system. Integrating this vision of the computing continuum with studies on causal inference allows us to leverage structural models and graph theories for extracting cause-effect relationships. This approach aids self-adaptation mechanisms for the computing continuum, making the overall system capable of adjusting to and generalizing in unexpected scenarios.

Such distributed systems are inherently complex; thus, developing management processes requires coordination of com-

2021 IEEE International Conference on Edge Computing (EDGE) | 978-1-6654-0062-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/EDGE53862.2021.00021

2767-9918/21/$31.00 ©2021 IEEE
DOI 10.1109/EDGE53862.2021.00021

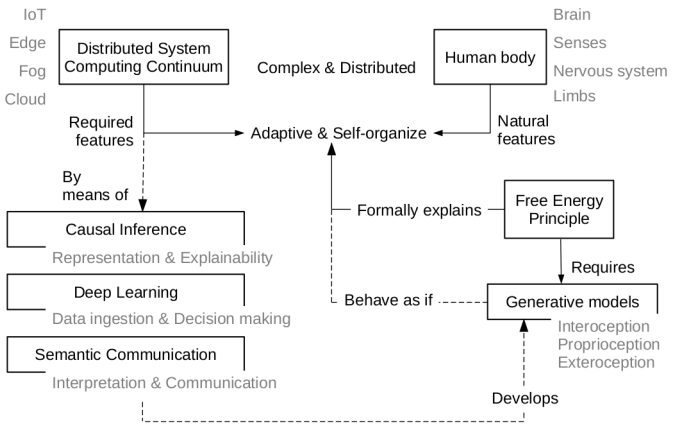

Fig. 1. General overview on the roadmap towards self-organizing and adaptive distributed systems of the computing continuum.

munication and actions. In this scenario, deep learning techniques can help manage the system's size, complexity, and data load, providing models and abstractions. In this direction, neuroscience-inspired methodologies, e.g., *shared global workspace*, can play an important role.

Finally, the distributed fashion of the system and the use of independent components call for the use of *semantic communication*, where the transmitted data within the parts also encode vital information such as significance or freshness of the message, but also, more specific information such as membership or relationships.

II. Background

The ever-growing complexity of distributed systems stems from their field of influence, no more isolated and partitioned, but currently spanning from a central, high-performance cloud to the smallest, lightweight sensor device. This scenario constitutes a complex ecosystem and brings many challenges: developing performative and interconnected functions, finding ways to communicate, and adapting to new technologies and ever-changing environments. Therefore, we can no more consider IoT, Edge, Fog, and cloud in isolation. Given these premises, traditional management methodologies become increasingly obsolete, falling short in providing the necessary solutions.

In this context, the capability of adaptation of the system becomes an essential requirement. Looking outside the engineering perspective and focusing on cognitive science, we can see how humans adjust to new, changing conditions, unlike current machines and systems. Thus, this section temporarily detours from the engineering field to highlight neuroscience theories on how the brain works, inspiring us to find methods and tools to describe and model this setting. In this process, we do not aim to achieve a one-to-one mapping but rather a translation work between still debated neuroscience theories and the realm of the computing continuum. Section II-A first describes high-level requirements for the computing continuum. Consequently, Sections II-B II-C, II-D provide background knowledge for neuroscience-related concepts.

A. Computing continuum management requirements

The complexity and scale of these systems challenge current methodologies for managing distributed Internet-based systems, and methodologies such as the elasticity for Cloud systems [9], [10], [11], and, most importantly, for online Cloud-Fog-Edge computing [12] are no longer fully applicable. Hence, every possible solution depends on the definition of novel, general, and adaptive requirements for computing continuum management (CCM).

1) Flexible representation: Due to its scale and variable composition, CCM needs a flexible and adaptive representation of the system. Given the mutable state of the system architecture, its representation shifts as a variable feature of the system.

2) Link with underlying infrastructure: The characteristics of computing continuum systems depend on their underlying infrastructure and technologies tightly coupling with the application. The variety of IoT devices, Edge, Fog, and Cloud configurations compels any CCM methodology to consider the infrastructure as a key component.

3) Causality relations: We can see the computing continuum as an ecosystem; thus, a global perspective cannot consider modules and their action in isolation. Therefore, CCM must keep track of *causal relationships* between its components to understand the consequences of modifying a part or trace back any possible system issue.

4) Temporal evolution: Real-time systems in a computing continuum constantly evolve due to the external environmental transformation and the intrinsic differentiation and extensibility characteristics. Their state and structure can change with time; hence, taking this temporal evolution into account in CCM provides long-term adaptation capabilities tools. We can consider the system's variables as evolutionary, thus extracting environmental derivatives to understand the "direction" that a system is taking and act proportionally.

5) Proactive adaptation: The complexity of the ecosystem can lead to a cascade of failures as issues can propagate due to components' interconnections. In this regard, the CCM needs to act and adapt as soon as possible to overcome the system's agitations before they propagate.

6) Learning framework: The ecosystem complexity and scale make it impossible to draw a complete management plan in the design phase. Therefore, setting management methodologies *inside* a learning framework is required to provide incrementally better solutions and adaptations.

7) Security and privacy embedded: It is an increasing concern *how* these systems can expose security and privacy. Therefore, any management methodology has to have embedded means to tackle these types of issues.

B. Free Energy Principle

A significant point emerging from the requirements specification in Section II-A is the inherent complexity of the

computing continuum ecosystem. In cognitive neuroscience, research proposed by Karl Friston intends to tackle the complexity of the brain, modeling it and its behavior according to what he calls the "Free Energy Principle" (FEP) [3]. At large, this principle states that, behaviorally, the brain tries to improve its model of the world by minimizing the Free Energy [1], i.e., the difference between the expected observation and the obtained one. This approach is captivating due to its mathematical foundation.

Mathematically, FEP's brain model leverages the concept of a Markov Blanket [13], a mathematical artifact in turn inspired by the work on reasoning and causal inference from Judea Pearl [14]. The Markov Blanket provides a framework for structuring causality on the system and formally separates it from its environment. We can see a Markov Blanket as a tool to infer one node, where the neighboring nodes encode all the knowledge needed to infer the target node. This method guarantees a wrapping tool for encoding knowledge since all the information needed to infer a node comes from its neighbors. As further contributions, this principle states that the system behavior is valid at different scales and that it is possible to construct hierarchical levels of Markov blankets [15]. In order to learn a more precise system representation (or model), i.e., to minimize the Free Energy, this principle proposes a methodology for a learning framework called active inference [16] (see Section II-C) that shares concepts with reinforcement learning.

Takeaway: With the premise that FEP is still under discussion [17], main ideas can be adopted to resolve the requirements for the computing continuum ecosystem. These features are relevant to developing top-down management methodologies capable of producing dynamical models, generalizing from the specific implementations.

C. Active inference and generative model

According to current cognitive neuroscience theories [18], different modules (regions) in the brain formulate hypotheses during reasoning. Each module then sends these *predictions* to the other modules, producing messages carrying probabilistic predictions of the external world. This approach is *top-down* since it starts from high-level areas of our brain and reaches the sensory areas. At this point, this top-down prediction combines with *bottom-up* messages from the external world, i.e., connect with the sensory signals. Here, the *generative model* computes a *prediction error*, i.e., the difference between what the interoceptive processes predicted and the observation. At this point, the system uses this error to update the internal model of the world. This generative model builds on top of three main elements, (i) a prediction, (ii) a prediction error, and (iii) the precision, i.e., the predictability of the signals. As defined by Clark in [19], predictive processing is the approach ascribed to our brain of predicting the sensory inputs with a generative model of the environment that best minimizes the external world perception. Active inference introduced in Section II-B is one of these theories.

In order to generate a model on such a complex ecosystem, there is the need to have a set of receptive functions to capture the ecosystem's temporal evolution [19]. Cognitive neuroscience theorizes that, in humans, we have three main sensory mechanisms to capture and represent energy perturbations. The first one is called *exteroception* and includes the outside-body stimuli and the connected sensory inputs. The other two channels are "inward-looking". One relates to *proprioception*, i.e., the sense of the relative position of the body in the environment; it connects to deployed actions (or forces). The last one, group is *interoception* or the sense of body "from within", i.e., all the signals internal to our body; it has a link to humans' perception [5]. Placing a predictive processing mechanism on top of these channels means creating a model that can better learn and engage with the ecosystem to minimize its representation error, i.e., the energy, and keep an *equilibrium*.

Takeaway: Considering the computing continuum as our ecosystem, we can thus build on top of it a generative model. The framework can interpret and proactively adapt to the signals coming from within our system and the environment deploying the proper strategies to maintain the overall equilibrium.

D. Global workspace theories

The last step is to understand how our brain treats the information that attends and chooses the actions to perform [18]. The answer seems to be in what cognitive neuroscience calls the *working memory*, i.e., a restricted space where we can keep in mind the relevant elements needed to solve a specific problem. Baars [20] and then Dehaene with an extended version [4] call this memory *Global Neuronal Workspace* (GNW), formulating what is currently one of the most corroborated sets of theories of consciousness [21]. We can put it in other terms: the shared workspace works as a *brain router* that can receive bottom-up and transmit top-down information from and towards the many processors in the brain [22]. The activation of the GNW happens in a non-linear way, called "ignition" [23]. The ignition characteristic is the exclusive activation of a subset of workspace neurons representing the conscious content code. According to this theory, this exclusively and capacity-limited architecture allows for the adaptive extraction of relevant information.

In this direction, this prospect attracted the attention of the deep learning community, seeing it as one of the potential influences for decision-making and generalizing learning towards AI [24]. VanRullen and Kanai [25] envisioned how deep learning techniques could represent the Global Workspace Theory, creating what they called a Global Latent Workspace. Recently, Goyal et al. [26] proposed a shared global workspace for the coordination of modules, testing it in multi-agent systems like Recurrent Independent Mechanisms, Transformers and Reinforcement Learning. Again, all of the techniques use the concept of *attention* [27] as a building block, i.e.,

[1]Friston's definition of Free Energy is different from the one commonly used in thermodynamics.

a method that through softmax functions can dynamically highlight components of the input to adapt to an ever-changing output.

Takeaway: Although there is an active debate on the connection of GNW and "consciousness" in philosophical and neuroscientific research [6], this theory draws captivating ideas. Its modeling could aid the engineering task of facilitating multiple specialized agents' communication in a complex environment and eventually drawing out causality dependencies, given the theorized capability of extracting higher-level definitions out of specialized processing modules.

III. Vision

Here, we aim at pragmatically presenting the methodologies introduced in Section II. We outline the setting for the management of the computing continuum, formulating its representation and the way to extract knowledge.

A. Delineate the computing continuum states

In order to manage computing continuum systems, the first step is to develop a *representation of its status*. This high-level representation has a threefold purpose. First, it has to provide *interpretability* to let all the system's stakeholders understand it. Second, this representation must be *unequivocal*. Moreover, third, it has to facilitate the link of the system with its underlying *infrastructure*. Given the listed requirements, the most appropriate high-level representation should be a composition of three main dimensions, "Resources", i.e., its usage and type, "Quality", i.e., the quantifiable system's performance and "Cost", i.e., the price of the system's configuration. Works in cloud elasticity used these dimensions to model the scenario in a *cartesian space*; however, they have some limitations on how these can represent the overall system and its temporal evolution in the computing continuum [28].

B. Infer the system's state

The limit of this high-level system representation is that it is not directly observable, and its latent representation can be the result of the hidden combination of lower-level variables. Probabilistic theories allow us to reason about latent variables given observable states. Graph representations tackle this task by helping build and guide a generative model of this ecosystem, i.e., the joint distribution of the observable and latent variables. The role of graphs in probabilistic reasoning is to (i) provide advantageous ways to express assumptions, (ii) to simplify the representation of joint distributions, and (iii) to facilitate inference starting from observations [29].

With the use of the Markov Blanket condition, it is possible to find a sufficient representation of the system. This Markov Blanket condition has the essential role of framing the inference problem in a tractable fashion by building structures that simplify the exponential problem of probabilistic reasoning under uncertainty. Under this condition, considering directed or undirected graphs, all the computations involving nodes of these graphs can occur in the local set of states represented by their neighborhood. The definition above drives us to model a graph of these three variables that leverages the Markov Blanket to infer them through this artifact. Hence, by taking advantage of observable system characteristics, it will be possible to infer the overall system state. Furthermore, in the case of Bayesian causal graphs (or Bayesian causal networks) [29], the Markov Blanket property works as a filter for the causality chain, i.e., only the neighboring nodes included in the Markov Blanket are necessary to explain the observed state without the need to reach the chain roots. Organizing knowledge with this layout allows a flexible representation of parent-child relationships. However, this modular configuration assumes that each relationship represents a stable and autonomous mechanism, i.e., that changing one relationship does not affect the others.

C. How to build the graph: the state's parents and children

Following the Markov Blanket representation of the brain proposed in [13], we decompose the neighborhood of a node into two groups, the *sensory states*, and the *action states*. The former represents the parents of the inferred node, and the latter the children.

Thinking of the global state of the system, we ask ourselves how to identify its parents. From the perspective of graph theory, this task implies finding the variables that *condition* the state. Determining these variables is not easy since they are diverse, and the same variables can acquire different meanings depending on the context. Thus, we state that the application requirements must define these specific variables and their correct relationship with the system state. In other words, we mean that some applications will require a set of variables different from others and that, in some situations, even if they share some variables, their relations to the central state can be different. Practically, this suggests that depending on the system requirements, Service Level Objectives (SLOs) [30], which can be defined as this set of variables, can have different relations to the central and higher-level system state representation. For example, a fleet management system and an autonomous car system can similarly have a variable that expresses latency; however, the first case can have more relaxed constraints, whereas the second crucially depends on that. Hence, given the large spectrum of applications on the computing continuum, for each of them, the set of sensor states and their relation with the high-level system state will need to be determined based on the system requirements.

Sensor states, which are required to be high-level abstractions, will also need to link with the system's underlying infrastructure, allowing the system to change its granularity and move towards lower-level abstractions whenever needed. Simply put, response time can be an aggregate and high-level representation of the system's latency. Then, it will be required to change its granularity to determine which specific part of the system is affecting its current value up to its underlying infrastructure.

The children of the high-level system state representation, also called active states in [13], are those nodes that are influenced

by the system state and can influence the environment. In this regard, these higher-level variables can act on the environment to adapt the system to it. Simply put, these can be seen as the variables that determine the system's configuration space; hence, their change affects how the system and its underlying infrastructure are related. It is important to remark that, even if these nodes are children of the central highest-level system state, they are required to infer it as they provide essential information about it.

D. Overall representation

At this point, we represent the system as a direct acyclic graph (DAG), with its central node being the system state. This DAG encodes causality relations from the metrics constructed from the system's underlying infrastructure up to the system state. Then, it links the system state with its adaptive capabilities given different configurations of the underlying infrastructure.

From this high-level perspective, a management structure for the computing continuum needs to understand how this information propagates through the system. Given that these are open systems and their environment is noisy, setting fixed thresholds and rules is limiting. Thus, we need to introduce the dynamic concept of equilibrium on the central state of the system. We can define that the equilibrium of such systems links a concise underlying infrastructure configuration, an operation mode, with the fulfillment of the application requirements. The process for achieving equilibrium requires defining the temporal evolution of the system. This process requires going beyond causality on the states, developing time derivatives of these causes. These derivatives will guarantee a long-term view of the system dynamics, differentiating short-term deviations from trends that eventually make the system lose its equilibrium. When the latter condition shows to be accurate, the system must act on the environment to achieve a new equilibrium state. This new state can be linked to a different underlying infrastructure configuration and, eventually, to a separate operation mode.

In addition, the knowledge of the system derivatives (or dynamics) allows better to analyze the impact of an action on the system. These systems are highly interconnected, which hinders deciding the proper action to avoid possible unexpected cascade effects. Hence, the system dynamics can help assess and limit the possible consequences of acting.

Figure 2 shows an overview of the entire system representation. At the left side of the figure, we can observe the computing continuum resources, the system's underlying infrastructure, and how it relates with the set of metrics; the environment influences those sensor states. Then, this set of metrics relates to the central state. It is important to remark here that the application requirements will sculpt these relations. Next, the central system's state influences the set of actions that can adapt the system to its underlying infrastructure. Finally, we can observe the system from two different perspectives, the first sets it in a temporal evolution, following the line on the bottom of the figure. The second is a learning perspective which inputs to the system the final output so that it can learn similarly as an agent in reinforcement learning.

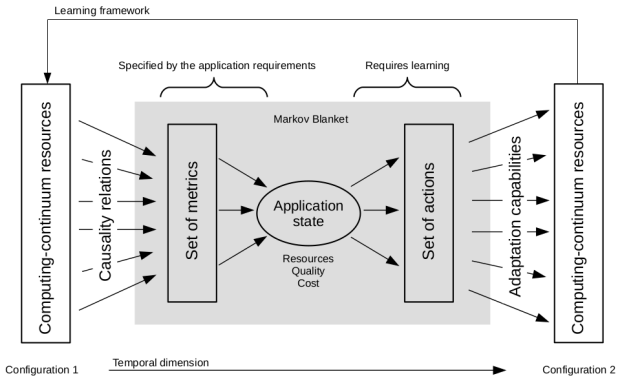

Fig. 2. Overall system representation

E. Estimate the model

The notion of equilibrium given above leads us to look for ways to define the most appropriate generative model for helping in policy decision-making. However, as described before, the signals are high-dimensional and noisy in real-life scenarios, and we do not always have the capability of generating the model; we just have the observed sequence of events. Thus, there is the need to find the best representation to contain the inherent information. In this case, developing a shared and restricted communication space, like the one theorized in the GNW, can help achieve this goal. According to the Free Energy Principle, the structure created through Markov blankets determines every action's salience. According to Nikolova et al. [31] the predictive processing theories like the GNW have a leading role in encoding and adjusting precision by resolving which predictions reach the conscious states. In practice, the ideas theorized by Hesp et al. in [32] and then formalized in the work of Whyte and Smith [33] develop the conceptualization of a predictive Global Neuronal Workspace. They define the predictive GNW as an active-inference-fueled process where the posterior expectation, i.e., the estimate of the posterior probability of generating the observed state, is performed through a softmax mechanism. This procedure guarantees an estimation of the precision, i.e., the capability of generating an interpretation of the sensory inputs given the internal processes.

This theory is relevant in describing the intrinsic self-model, i.e., the capability to self-inference the internal system organization. This approach aids the capability to generalize, starting from and modifying our perceptions. In the field of the computing continuum, this formula represents a way to create a higher-level representation of the systems' internal modules, i.e., the independent mechanisms in the graph, which constitute the basis for the active inference. Plus, using a restricted, softmax-based approach, the system is aware

of the precision of the current perception, thus acting on it. This memory-restricted representation of the observed space is coherent with the idea of a sparse Markov blanket, where only a few variables co-occur at the same time during active inference [13], [17]. In practice, graph-based frameworks like the one proposed by Huang et al. [34] go in this direction. In this case, the authors use a Variational Autoencoder to characterize a minimal set of state representations that serve for policy learning.

F. Ways to communicate

Furthermore, as in the neuroscience context of the GNW, "ignition" links to conscious cognition, which is verbalizable; consequently, this restricted attention-based memory can potentially help create a communication scheme for a complex system built of distinct agents [35], [36]. Indeed, this encoding approach aligns with the idea of discrete *semantic communication* introduced by Shannon [37], which relates the data transmission to the statistical probabilities connected to the event that generated it. Researchers and practitioners are working towards defining semantic-aware communication for agents (or mechanisms) in complex systems like the computing continuum [38]. In this context, Seo et al. [39] propose a stochastic communication method where agents can self-reason on creating the most appropriate semantic coding. Similarly, Kountouris and Pappas [40] discuss semantic-aware networking, presenting necessary characteristics like the capability of filtering, preprocessing, reception, and control. In this direction, modern Machine learning techniques, like [41], can be studied to address this problem. Overall, these works go in the similar direction of using some sort of encoding mechanisms to express and communicate a high-level representation of events and actions, indicating the potential of the proposed approaches.

IV. Discussion and future projections

This paper presents a direction towards the management of the computing continuum. In our vision, developing a hierarchical, graph-based representation of the system's metrics, requirements, actions, and states guarantees the generalizability of the cloud continuum. This goal is achievable through the definition of high-level representations and self-organization over the temporal evolution of the system.

In this context, we incorporate neuroscience theories extracting methodologies for achieving generalization and equilibrium within the system. This set of neuroscience-based hypotheses and postulates help to the extent that they guarantee the interpretation of signals and a proactive generation of learning models to interpret and act in the environment. We are mindful of the fact that what we propose is not an attempt of mapping human-like intelligence in the computing continuum scenario, a goal that goes beyond our scope, and we are aware of the risks of "impliciting" human-machines metaphors in the field of computer science [42].

Given these premises, our vision opens up challenges and future projections to achieve a more cohesive and self-adaptive representation of the computing continuum. An essential step is defining ways to express the prior knowledge [43], which we can see in the form of prior beliefs [44], [15] or inductive biases [36], needed to represent a causal model of the system we are trying to shape. This stage sets the basis for the definition of the applications' graph; thus, it is fundamental that the translation of requirements and actions are accurate.

Another critical task is the integration of the causal model of the computing continuum with its inference mechanisms. The techniques and approaches described in Section III-E go in this direction, but there is the need to test them on a large-scale scenario.

Finally, we require to have semantic representation of high-level concepts to guarantee a shared communication language. Furthermore, we want this communication to be compute-continuum-wide, i.e., able to connect and combine signals coming from the system's different (geographical and logical) regions, a stream of works and concepts, presented in Section III-F address this topic. However, there is the necessity to inspect the generalization and performance capabilities of these methodologies.

V. Conclusion

To handle the uncertainty and complexity of distributed systems, we want to move forward from single-tier management approaches and consider solutions that work on the whole computing continuum, i.e., from the Edge to the Cloud of Internet systems. In this context, the interconnection of devices and systems and the conditioning deriving from their actions are so rooted in a way that makes using rule-based local management solutions insufficient. In this paper, we attempted to clarify how, examining solutions that consider generalizability and high-level representations as their target, we can build systems able to better self-organize over time in changing environments. Much effort is, of course, still needed to connect the various methods and extract suitable ways to incorporate this proposal in computing continuum frameworks, but we set the basis for a new stream of research.

References

[1] P. Maciel, J. Dantas, C. Melo, P. Pereira, F. Oliveira, J. Araujo, and R. Matos, "A survey on reliability and availability modeling of edge, fog, and cloud computing," *Journal of Reliable Intelligent Environments*, pp. 1–19, 2021.

[2] S. Dustdar, O. Mutlu, and N. Vijaykumar, "Rethinking Divide and Conquer-Towards Holistic Interfaces of the Computing Stack," *IEEE Internet Computing*, vol. 24, no. 6, pp. 45–57, nov 2020.

[3] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *Journal of Physiology Paris*, vol. 100, no. 1-3, pp. 70–87, jul 2006.

[4] S. Dehaene and L. Naccache, "Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework," *Cognition*, vol. 79, pp. 1–37, 2001.

[5] A. K. Seth and K. J. Friston, "Active interoceptive inference and the emotional brain," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1708, p. 20160007, 2016.

[6] O. Carter, J. Hohwy, J. Van Boxtel, V. Lamme, N. Block, C. Koch, N. Tsuchiya *et al.*, "Conscious machines: Defining questions," *Science*, vol. 359, no. 6374, pp. 400–400, 2018.

[7] V. Casamayor Pujol, P. Raith, and S. Dustdar, "Towards a new paradigm for managing computing continuum applications," in *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2021, to appear.

[8] J. Pearl, "Causal inference," in *NIPS Causality: Objectives and Assessment*, 2010.

[9] S. Dustdar, Y. Guo, B. Satzger, and H. L. Truong, "Principles of elastic processes," *IEEE Internet Computing*, vol. 15, no. 5, pp. 66–71, sep 2011.

[10] P. Hoenisch, D. Schuller, S. Schulte, C. Hochreiner, and S. Dustdar, "Optimization of Complex Elastic Processes," *IEEE Transactions on Services Computing*, vol. 9, no. 5, pp. 700–713, sep 2016.

[11] S. Dustdar, V. Casamayor Pujol, and P. Kumar Donta, "On distributed computing continuum systems," *IEEE Transactions on Knowledge and Data Engineering*, 2021, to appear.

[12] M. D. de Assunção, A. da Silva Veith, and R. Buyya, "Distributed data stream processing and edge computing: A survey on resource elasticity and future directions," *J. Netw. Comput. Appl.*, vol. 103, pp. 1–17, 2018.

[13] I. Hipolito, M. Ramstead, L. Convertino, A. Bhat, K. Friston, and T. Parr, "Markov Blankets in the Brain," *Neuroscience and Biobehavioral Reviews*, vol. 125, pp. 88–97, jun 2020. [Online]. Available: https://arxiv.org/abs/2006.02741v1

[14] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[15] E. R. Palacios, A. Razi, T. Parr, M. Kirchhoff, and K. Friston, "On Markov blankets and hierarchical self-organisation," *Journal of Theoretical Biology*, vol. 486, p. 110089, feb 2020.

[16] N. Sajid, P. J. Ball, T. Parr, and K. J. Friston, "Active Inference: Demystified and Compared," *Neural Computation*, vol. 33, no. 3, pp. 674–712, mar 2021. [Online]. Available: https://doi.org/10.1162/neco_a_01357

[17] V. Raja, D. Valluri, E. Baggs, A. Chemero, and M. L. Anderson, "The Markov blanket trick: On the scope of the free energy principle and active inference," *Physics of Life Reviews*, sep 2021.

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

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

[20] B. J. Baars, *A cognitive theory of consciousness*. Cambridge University Press, 1993.

[21] M. Michel, S. M. Fleming, H. Lau, A. L. Lee, S. Martinez-Conde, R. E. Passingham, M. A. Peters, D. Rahnev, C. Sergent, and K. Liu, "An informal internet survey on the current state of consciousness science," *Frontiers in psychology*, vol. 9, p. 2134, 2018.

[22] G. A. Mashour, P. R. Roelfsema, J. P. Changeux, and S. Dehaene, "Conscious processing and the global neuronal workspace hypothesis," *Neuron*, vol. 105, pp. 776–798, 2020.

[23] S. Dehaene, C. Sergent, and J. P. Changeux, "A neuronal network model linking subjective reports and objective physiological data during conscious perception," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 8520 – 8525, 2003.

[24] Y. Bengio, Y. LeCun, and G. E. Hinton, "Deep learning for ai," *Communications of the ACM*, vol. 64, pp. 58 – 65, 2021.

[25] R. van Rullen and R. Kanai, "Deep learning and the global workspace theory," *Trends in Neurosciences*, vol. 44, pp. 692–704, 2021.

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

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2015.

[28] H.-L. Truong, S. Dustdar, and F. Leymann, "Towards the realization of multi-dimensional elasticity for distributed cloud systems," *Procedia Computer Science*, vol. 97, pp. 14–23, 2016.

[29] J. Pearl *et al.*, "Models, reasoning and inference," *Cambridge, UK: CambridgeUniversityPress*, vol. 19, 2000.

[30] S. Nastic, A. Morichetta, T. Pusztai, S. Dustdar, X. Ding, D. Vij, and Y. Xiong, "SLOC: Service level objectives for next generation cloud computing," *IEEE Internet Computing*, vol. 24, no. 3, pp. 39–50, may 2020.

[31] N. Nikolova, P. T. Waade, K. J. Friston, and M. Allen, "What might interoceptive inference reveal about consciousness?" *Review of Philosophy and Psychology*, 2021.

[32] C. Hesp, R. Smith, T. Parr, M. Allen, K. J. Friston, and M. J. D. Ramstead, "Deeply felt affect: The emergence of valence in deep active inference," *Neural Computation*, vol. 33, pp. 398 – 446, 2021.

[33] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301008220301738

[34] B. Huang, C. Lu, L. Leqi, J. M. Hernández-Lobato, C. Glymour, B. Schölkopf, and K. Zhang, "Action-sufficient state representation learning for control with structural constraints," *arXiv preprint arXiv:2110.05721*, 2021.

[35] Y. Bengio, "The consciousness prior," *arXiv preprint arXiv:1709.08568*, 2017.

[36] A. Goyal and Y. Bengio, "Inductive biases for deep learning of higher-level cognition," *arXiv preprint arXiv:2011.15091*, 2020.

[37] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[38] E. Uysal, O. Kaya, A. Ephremides, J. Gross, M. Codreanu, P. Popovski, M. Assaad, G. Liva, A. Munari, T. Soleymani *et al.*, "Semantic communications in networked systems," *arXiv preprint arXiv:2103.05391*, 2021.

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

[40] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 96–102, 2021.

[41] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *arXiv preprint arXiv:2103.03230*, 2021.

[42] A. T. Baria and K. Cross, "The brain is a computer is a brain: neuroscience's internal debate and the social significance of the computational metaphor," *arXiv preprint arXiv:2107.14042*, 2021.

[43] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard, "On pearl's hierarchy and the foundations of causal inference," 2020.

[44] M. Kirchhoff, T. Parr, E. Palacios, K. Friston, and J. Kiverstein, "The markov blankets of life: autonomy, active inference and the free energy principle," *Journal of The royal society interface*, vol. 15, no. 138, p. 20170792, 2018.