Intelligent Service Adaptations through Active Inference Agents

Boris Sedlak, Victor Casamayor Pujol, Praveen Kumar Donta, and Schahram Dustdar

Distributed Systems Group, TU Wien, Vienna 1040, Austria.
{b.sedlak, v.casamayor, pdonta, dustdar}@dsg.tuwien.ac.at

Introduction

Large-scale distributed systems can be composed of multiple components and computing tiers, all of which have unique contributions to the system's high-level objectives. Consider, for instance, that videos captured by surveillance cameras can be processed on nearby devices and then streamed over the internet. To ensure the system's functionality, stakeholders describe each tier's expected behavior through Service Level Objectives (SLOs), e.g., maintaining processing latency under a certain boundary. Evaluating these SLOs requires a set of metrics (i.e., sensory observations), which are commonly collected at one central system location; given this data, it can then be calculated to what degree SLOs were fulfilled. However, this requires transferring massive amounts of data; further, the latency for detecting and resolving SLO violations is high. The human body, as an example of a complex system, would collapse from the overhead of evaluating each cell's requirements centrally (i.e., in the brain); hence, requirements assurance must be decentralized to the respective system components.

Given that its SLOs are violated, intelligent components should detect and resolve this autonomously by interacting with their environment. Understanding how to adapt a system requires in-depth knowledge, which can be provided through Machine Learning (ML) techniques. In particular, this can involve extracting causal relationships between components. However, training such models requires large amounts of data upfront, which are also subject to variable shifts over time. This promotes the usage of Active Inference (AIF) for two reasons: (1) AIF provides continuous model accuracy while creating world models without prior assumptions, and (2) AIF agents aim to persist over time, thus they can be used to modify the system according to expected SLO fulfillment and model improvement. Within our most recent works [1,2,3] we addressed these matters; the following presents a condensed version of our contributions.

Methodology

Figure 1 shows how individual system components gradually develop a causal understanding of how to ensure their SLO. For this, consider a video processing workload (right); during the execution, metrics are extracted as part of the processing environment. In parallel, an AIF agent predicts whether it expects SLOs (e.g., $latency < t$) to be fulfilled under the given environment (left); then, the agent compares the prediction with the actual observations and updates its beliefs according to the surprise, i.e., minimizing the free energy. For the agent, this means updating its underlying causal graph and the encoded conditional probabilities. Afterward, the agent compares the probability of SLO violations under different environmental states and suggests changes to the processing environment, e.g., adapting the video streaming parameters. To narrow down which variables have a direct causal impact on SLO fulfillment, we extract the Markov blanket (MB) around SLO variables; consider for this Figure 2a: given an SLO that aims to cap latency (green), adjusting its parent variables (blue) directly affects *latency*, while the remaining variables (grey) can be disregarded.

In this example, an AIF agent executed at the processing device (see Figure 1) might use its local scope to adjust the camera's video resolution (*pixel*), whereas offloading *streams* between agents would require a higher-level orchestrator to communicate between agents. This is achieved through collaboration between individual components, which raises the system-wide level of intelligence. To accelerate this, agents exchange their generative models according to the relative differences between their processing environments. As lower-level tiers get spanned with intelligent components, higher-level tiers can rely on their correct function when they construct services on top of them. Such mechanisms are also applied by the human body, which assembles higher-level components (e.g., muscles or organs) from smaller cellular structures; thus, the equilibrium within each cell contributes its part to the system's global objectives.

Evaluation

The presented framework was implemented[1]and evaluated for the given use case, in which a distributed system is responsible for ensuring service-related SLOs within a distributed video streaming architecture. The evaluation included a total number of twelve aspects, such as the number of training rounds to converge to satisfying SLO fulfillment or the extent to which an MB can decrease the complexity of inference. Figure 2b depicts the SLO fulfillment based on how an AIF agent adapts its environment; within its local scope, the agent could adjust the *pixel* and frames per second (*fps*) of the video stream. Whenever the AIF agent decided to switch to another configuration (blue dots), this showed a decisive impact on the SLO fulfillment; in total, it required 5 reconfigurations and 16 respective AIF iterations for the function to converge.

Another important property of the solution is that the causal structures were rationally explainable, which improves trustworthiness. As in human interaction, the possibility to reason about decisions, and reflect why a configuration was taken under some circumstances, proves crucial to foster understanding. By exchanging this knowledge between agents, they contribute to a general world model, which allows them to accurately adapt their environments.

[1] Prototype artifact available at GitHub, accessed May 22nd 2024

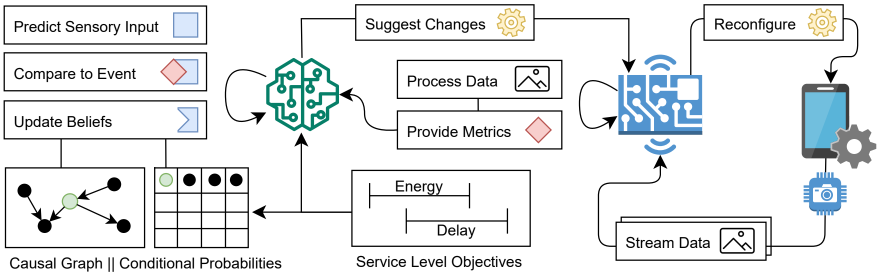
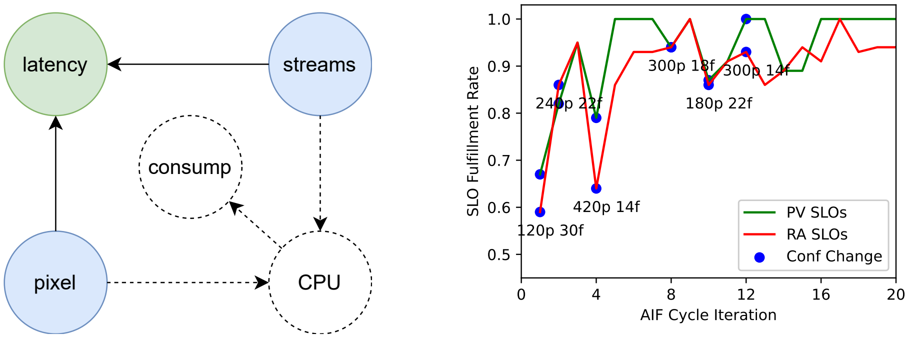


Fig. 1: High-level AIF implementation comprising perception and action



(a) MB extracted around *latency*

(b) SLO fulfillment under AIF agent

Fig. 2: Inferring SLO-compliant configurations based on Markov blankets

References

1. Dustdar, S., Pujol, V.C., Donta, P.K.: On Distributed Computing Continuum Systems. IEEE Transactions on Knowledge and Data Engineering **35**(4), 4092–4105 (Apr 2023). https://doi.org/10.1109/TKDE.2022.3142856
2. Sedlak, B., Pujol, V.C., Donta, P.K., Dustdar, S.: Equilibrium in the Computing Continuum through Active Inference (Nov 2023). https://doi.org/10.48550/arXiv.2311.16769, (Under Revision at FGCS)
3. Sedlak, B., Pujol, V.C., Donta, P.K., Dustdar, S.: Active Inference on the Edge: A Design Study. In: 2024 IEEE PerCom Workshops. pp. 550–555 (Mar 2024). https://doi.org/10.1109/PerComWorkshops59983.2024.10502828