X-by-Construction methods for scalable and trustworthy Cloud-Edge-IoT networks

Expression of Interest

Jürgen Becker[1], Tobias Dörr[1], Nikolaos Voros[2]
[1]Karlsruhe Institute of Technology, Germany
[2]University of the Peloponnese, Greece

on behalf of the XANDAR consortium
xandar-project.eu

1 MOTIVATION

A computing continuum that combines resources from the cloud, the edge, and Internet-of-Things (IoT) devices promises various benefits. Computation offloading for the purposes of energy efficiency, for example, is currently the subject of various studies.[1] In addition, a seamless connectivity between nodes in this continuum can improve customer experience, solve intellectual property issues, and simplify the deployment of software updates.

From an engineering perspective, the design of scalable and trustworthy cloud-edge-IoT systems is a challenging task. Open architectures are an enabler for a scalable network of federated infrastructures, but the large number of involved stakeholders makes it difficult to build them without targeted coordination and standardisation activities.

Furthermore, from a trustworthiness point of view, non-functional constraints such as functional safety requirements or the need to ensure the confidentiality of sensitive data are a barrier in the adoption of this continuum. Especially in cyber-physical system domains (healthcare, transportation, energy sector, …), failures along the cloud-edge-IoT chain can result in serious consequences all the way to physical harm or the loss of human life.

To overcome such barriers, future technologies must provide standardised network management functions and establish seamless connectivity between heterogeneous devices in the continuum. Time-predictable components with fault detection and fault tolerance support must be integrated in a reliable manner. With respect to artificial intelligence (AI) applications, energy-efficient accelerators with a low latency need to be complemented with multi-tenant accelerators for special purposes and applications. To optimise relevant properties during runtime, models to estimate the power-performance-area (PPA) metric of jobs, such as the execution of a computationally intensive AI algorithm, are necessary.

2 CURRENT STATUS

Supporting the federation across multiple clouds has already been the subject of several large-scale efforts, including European ones. However, extending the federation problem to the scope of edge and IoT nodes comes with a plethora of additional challenges. There is extreme platform heterogeneity in terms of hardware architectures, which often manifests itself in processing, memory, and data storage capabilities that differ by orders of magnitude. Furthermore, heterogeneous communication technologies are characterised by diverse bandwidth, communication delay, power consumption, and robustness properties.

Implementing the necessary technologies is further complicated by the inherently adaptive nature of the computing continuum. Traditional solutions to achieve time predictability and fault tolerance, for example, are often based on a static system architecture known during design time. Novel approaches such as the reactor-based coordination language Lingua Franca[2] are promising solutions to achieve a shift from static to dynamic execution models, but their integration into a large-scale cloud-edge-IoT network is yet to be explored.

[1] X. Li, Y. Dang, M. Aazam, X. Peng, T. Chen, and C. Chen, “Energy-Efficient Computation Offloading in Vehicular Edge Cloud Computing,” in IEEE Access, vol. 8, pp. 37632-37644, 2020, doi: 10.1109/ACCESS.2020.2975310.
[2] M. Lohstroh, C. Menard, S. Bateni, and E.A. Lee. “Toward a Lingua Franca for Deterministic Concurrent Systems,” in ACM Trans. Embed. Comput. Syst., vol. 20, issue 4, 2021, doi: 10.1145/3448128.

The novel X-by-Construction (XbC) approach[3] is defined as the *"step-wise refinement process from specification to code that automatically generates software (system) implementations that by construction satisfy specific non-functional properties."* The goal of the ongoing XANDAR project[4] is to develop an XbC-based toolchain that auto-generates embedded system implementations with guaranteed safety, security, and real-time properties from a model-based specification. Due to its focus on static systems, the current version of the XANDAR toolchain has limited applicability to an adaptive computing continuum. Nevertheless, its XbC-based structure is a promising platform to develop the necessary technology stack.

3 RESEARCH CHALLENGES

The main question to address is how existing architectures for the federation of European cloud ecosystems can be extended to connect the heterogeneous device landscape into a seamless cloud-edge-IoT federation. In this federation, edge and IoT devices need to become part of the value chains in a computing continuum. For the "Next Generation of IoT" to succeed, standardised solutions for the efficient utilisation of accelerators (such as GPUs, FPGAs, NPUs) on resource constrained multi-tenant devices will have to be developed.

As technical foundation to achieve this goal, the XbC paradigm is a promising platform. It will be fruitful to research approaches that auto-generate components for the computing continuum and deploy these components automatically. Therefore, the XbC paradigm must be adapted to make it compatible with the inherently adaptive nature of the envisaged ecosystem. A possible research direction is to integrate it with the existing DevOps flow (Figure 1). Conventionally, XbC procedures are executed only during the "build" phase of the system. By running them also during the release, the deployment, and the operation of components, dynamic properties (such as environment factors or variable hardware properties) can incorporated. Therefore, a distributed architecture of "XbC nodes" running generation and validation procedures (Figure 2) is a promising idea. These procedures could, for example, (1) predictively request a pair of edge nodes to act as a redundant pair of execution units for fault tolerance purposes or (2) adapt an Operational Design Domain (ODD) monitoring module whenever a self-learning AI algorithm incorporates new training data.

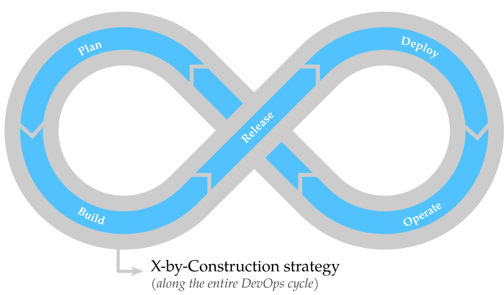

Figure 1: DevOps cycle integrated into an XbC concept

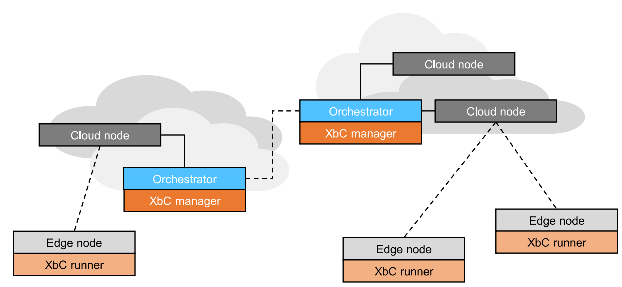

Figure 2: Cloud-edge network with distributed XbC nodes

For the successful implementation of this concept, various low-level issues need to be addressed. AI accelerators and GPUs, for example, are throughput-optimised architectures and therefore ill-suited for real-time applications. The problem becomes even more challenging when multi-tenant operation in needed. Low-level solutions to such challenges must be developed and can then be fed to the XbC concept as an input.

With a library of suitable low-level mechanisms and an adaptive XbC runtime that combines these mechanisms to ensure that relevant requirements are met in all operational phases, the problem to achieve a scalable and trustworthy cloud continuum can be structured and tackled.

[3] M.H. ter Beek, L. Cleophas, I. Schaefer, and B.W. Watson. "X-by-Construction," in Leveraging Applications of Formal Methods, Verification and Validation: Modeling (ISoLA 2018), Lecture Notes in Computer Science, vol 11244, 2018, Springer, Cham.

[4] L. Masing et al., "XANDAR: Exploiting the X-by-Construction Paradigm in Model-based Development of Safety-critical Systems," 2022 Design, Automation & Test in Europe Conference & Exhibition (DATE), Antwerp, Belgium, 2022, doi: 10.23919/DATE54114.2022.9774534.