

Developing Self-Adaptive Microservice Systems: Challenges and Directions

Nabor C. Mendonça, Pooyan Jamshidi, David Garlan, and Claus Pahl

Abstract—A self-adaptive system can dynamically monitor and adapt its behavior to preserve or enhance its quality attributes under uncertain operating conditions. This article identifies key challenges for the development of microservice applications as self-adaptive systems, using a cloud-based intelligent video surveillance application as a motivating example. It also suggests potential new directions for addressing most of the identified challenges by leveraging existing microservice practices and technologies.

Index Terms—self-adaptive systems, microservices, DevOps, continuous delivery.

1 INTRODUCTION

A self-adaptive system can monitor its behavior and change its configuration or architecture at run time to preserve or enhance its quality attributes (e.g., performance, reliability, and security) under uncertain operating conditions (e.g., varying workloads, errors, and security threats) [1]. Despite significant progress in recent years in developing self-adaptive systems, only a handful of techniques, such as automated server management, cloud elasticity, and automated data center management, have thus far found their way to industrial applications [2]. For example, Kubernetes,¹ a modern container orchestration platform that is increasingly being used to deploy and manage self-adaptive microservice applications in the cloud [3], only provides autoscaling, i.e., the ability to automatically change the number of instances of a service, and self-healing, i.e., the ability to automatically restart failed service instances, as part of its native self-adaptation capabilities.

The popularity of microservices in industry has naturally sparked the interest of the research community. However, most existing research on microservices focus on general architectural principles and migration guidelines (e.g., [4]). Only a few works have addressed the specific challenges of developing microservice applications as self-adaptive systems, but even those tend to be rather narrow in scope, offering only limited forms of adaptation (e.g., self-healing [5] and run-time placement adaptation [6]).

We believe the uptake of microservices and their enabling practices and technologies by industry brings a

unique opportunity to narrow the gap between the state-of-the-art and the state-of-the-practice in self-adaptive systems, and that both self-adaptive systems and microservice communities have much to gain from each other.

On the one hand, recent progress in self-adaptive systems offers a control-oriented perspective that leverages a large body of theoretical and practical results to enhance microservice quality attributes using techniques such as planners (e.g., to select the best possible adaptation strategy for each microservice), machine learning (e.g., to learn new adaptation strategies from past adaptation results), reasoning under uncertainty (e.g., to cope with noisy monitoring data), and multi-objective optimization (e.g., to cater for multiple, possibly conflicting microservice requirements) [7]. On the other hand, software characteristics such as independent and frequent deployment, the need for a high degree of automation, and complex run-time architectures [3] make microservices a fertile ground to foster further research and development on self-adaptive systems.

In this paper, we take a closer look into the interplay between self-adaptive systems and microservices in order to identify key challenges for the development of microservice applications as self-adaptive systems. Our contributions are as follows: (1) we describe a cloud-based intelligent video surveillance application as an example of a self-adaptive microservice system; (2) we identify and illustrate, in the context of this example application, several challenges for microservice development, delivery and operations from multiple self-adaptation perspectives; and (3) we discuss potential new directions for addressing most of the identified challenges by leveraging existing microservice practices and technologies.

2 MICROSERVICES

Microservices constitute an emerging architectural style that builds on the well-established concept of modularization but emphasizes technical boundaries (i.e., different address and execution spaces) between software components [3]. Each module—each microservice—is developed around a single business capability, offering access to its internal logic and data through a well-defined network interface. Due

- N. C. Mendonça is with the Post Graduate Program in Applied Informatics, University of Fortaleza, Fortaleza, CE, Brazil.
E-mail: nabor@unifor.br
- P. Jamshidi is with the Computer Science and Engineering Department, University of South Carolina, Columbia, SC, USA.
E-mail: pjamshid@cse.sc.edu
- D. Garlan is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.
E-mail: garlan@cs.cmu.edu
- C. Pahl is with the Faculty of Computer Science, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy.
E-mail: cpahl@unibz.it

1. <https://kubernetes.io/>

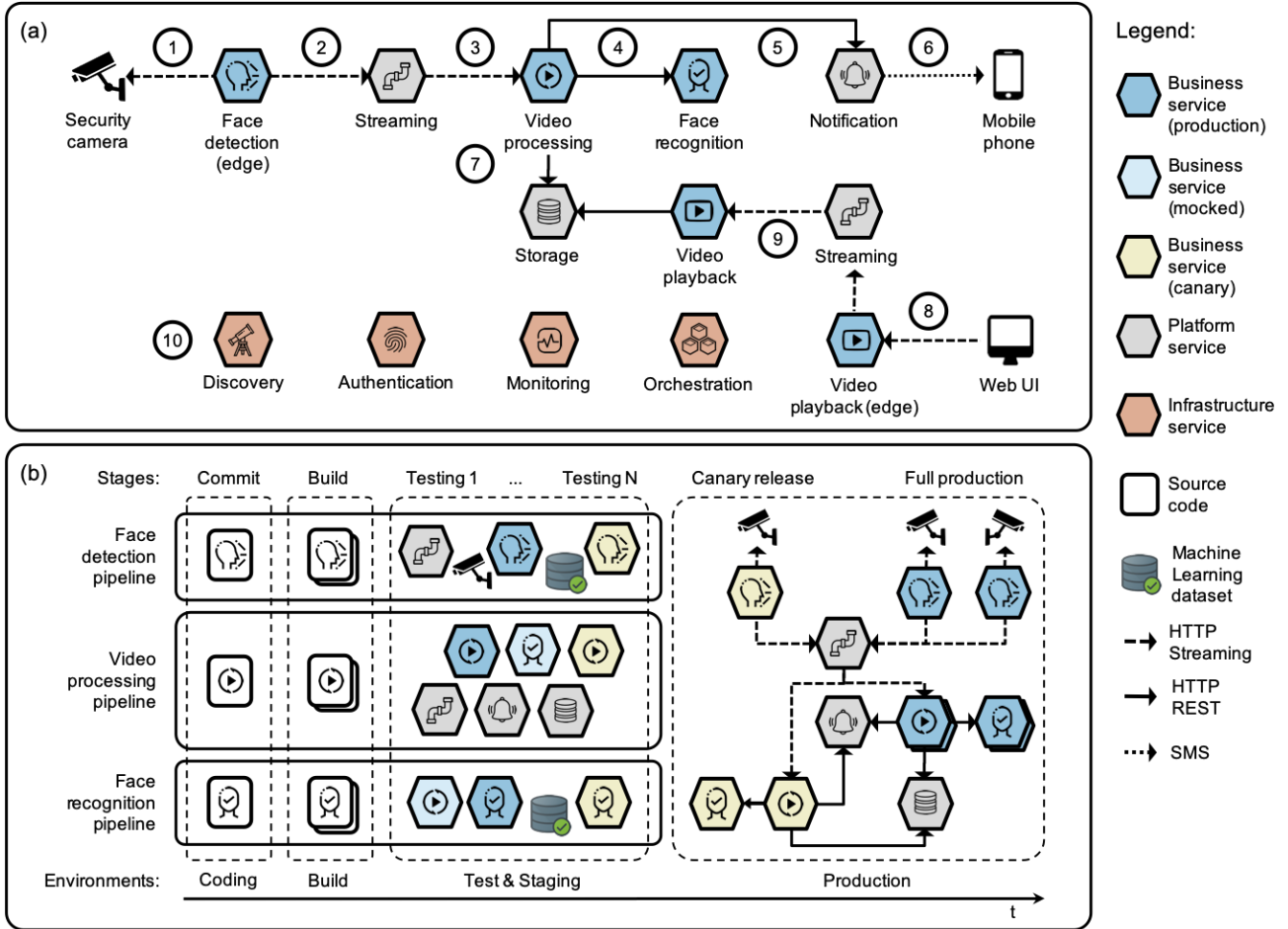


Fig. 1. (a) Microservice architecture for the edge-cloud intelligent video surveillance application; and (b) three continuous delivery pipelines for the application's face detection, video processing, and face recognition microservices, respectively (adapted from [8]).

to their relative simplicity and small size, microservices can be released often and adapted to different production environments, from cloud data centers to resource-constrained devices (e.g., IoT), without much development effort. This has a significant impact on improving software agility, because each microservice becomes an independent unit of development, deployment, versioning, scaling and management [3].

Microservices are considered to be an enabler for emerging DevOps practices, such as continuous integration (CI) and continuous delivery (CD), which aim to significantly decrease the time between changing a system and transferring that change to the production environment [9]. To achieve this level of agility, microservices are typically packaged in lightweight containers, e.g., Docker,² and deployed and managed using automated container orchestration tools, e.g., Kubernetes. In particular, the use of fully automated CD tools, e.g., Spinnaker,³ enables the creation of independent CD pipelines for each microservice. Having multiple CD pipelines running in parallel allows microservice changes to be rapidly delivered into production, since only the affected microservice needs to be built and tested [8].

3 A SELF-ADAPTIVE MICROSERVICE SYSTEM

We consider an edge-cloud intelligent video surveillance application as an example of a self-adaptive microservice system. The application's goal is to alert users about the presence and possibly identification of humans in a certain location via real-time analysis of video frames captured by one or more security cameras. Its architecture, inspired by one of Amazon's machine learning (ML) sample solutions,⁴ is composed of several *business*, *platform*, and *infrastructure* services, as shown in Figure 1(a).

Each security camera is deployed along with a ML-based *face detection* edge service.^① This service is a less accurate yet more energy efficient version of a ML-based *face recognition* service deployed in the cloud. The role of the face detection edge service is to select relevant video frames (i.e., frames with detected human faces) to be asynchronously transmitted to a *video processing* service in the cloud using a cloud-provided *streaming* service.^② The streaming service offers a more scalable and reliable solution over the basic HTTP Streaming protocol. The video processing service in the cloud analyzes the received video frames^③ and passes them as invocation parameters to the face recognition

2. <https://www.docker.com/>

3. <https://www.spinnaker.io/>

4. <https://aws.amazon.com/blogs/machine-learning/create-a-serverless-solution-for-video-frame-analysis-and-alerting/>

service,⁴ in an attempt to recognize any person that might be visible in them. If the face recognition service reports a match, the video processing service sends out an SMS message informing about the match to the application users' mobile phones, using a notification service.^{5,6} The video processing service then saves the analyzed video frames along with any relevant information regarding the video recording (e.g., date, camera ID) and contents (e.g., names of the recognized persons, match accuracy) in a storage service.⁷ Application users can use a Web-based User Interface (Web UI) to search for and play back any analyzed video segment stored in the cloud. This is done by invoking a video playback edge service deployed somewhere closer to the user's physical location.⁸ The video playback edge service, in turn, uses the streaming service to communicate asynchronously with a video playback service in the cloud.⁹

The video surveillance application also relies on several infrastructure services, such as a discovery service, an authentication service, a monitoring service, and a container orchestration service.¹⁰ To avoid cluttering, interactions between business services and infrastructure services have been omitted from Figure 1(a).

Following typical microservice practices, each service of the video surveillance application is delivered in production using its own independent CD pipeline. Figure 1(b) depicts three possible CD pipelines for the face detection, video processing, and face recognition services, respectively. Note how each pipeline goes through different environments, from coding to production, representing the multiple development stages (e.g., build, testing, canary release, full production) of each service [9].

We envision multiple self-adaptation scenarios for this example application. For instance:

- The video processing service may need to dynamically change its number of deployed instances in response to load variations;
- The video recognition service may need to dynamically change its container image, e.g., to switch to a less accurate version, under extreme load conditions;
- The video playback service may need to dynamically change its quality attributes, e.g., frame rate, to cope with latency fluctuations; and
- The CD pipeline for the video playback edge service may need to dynamically adjust that service's testing parameters, e.g., to expedite testing of its self-healing capabilities during staging.

Enacting all those different scenarios may bring up a number of challenges for microservice application developers, as we discuss next.

4 CHALLENGES

We identify the main challenges facing the development of self-adaptive microservice systems from four perspectives: *design space*; *control loop deployment*; *continuous delivery*; and *testing*.

4.1 Design Space

Designing self-adaptive systems involves making design decisions about observing the environment and the system

itself, selecting adaptation mechanisms, and enacting those mechanisms [2]. In the context of a microservice application, the design space for making self-adaptation decisions is even more complex, due to the large number of run-time components and their independent and highly dynamic nature [3]. Thus, a first challenge is:

[C1] How to determine monitoring and adaptation mechanisms to face the diversity of microservices' quality attributes.

In the case of the example video surveillance application, for instance, the quality requirements and adaptation needs of the ML face recognition service, which handles only individual video frames and is invoked synchronously from within the cloud, might be quite different from those of the video playback edge service, which handles the entire video segments and is invoked asynchronously from outside the cloud. In particular, the former would not have to monitor and adapt its communication parameters (e.g., request rate) at run time, which would be a critical requirement for the latter.

Another challenge is:

[C2] How to identify and resolve potential conflicts between the quality requirements of individual microservices and those of the overall application when defining their self-adaptive behaviors.

For instance, an important quality attribute of a face recognition service, which typically works by comparing extracted facial features from a given image with faces within a database, is its recognition accuracy. However, high accuracy in this kind of ML service inevitably implies high processing and storage costs, which might conflict with the application's overall cost constraints.

A further challenge in this context is:

[C3] How to reconcile the adaptation needs of individual microservices and the overall application with the self-adaptation capabilities offered by the underlying infrastructure management platform.

For example, while Kubernetes' autoscaling capabilities could be useful to improve the response time of all cloud-based services deployed as part of the video surveillance application, current Software Reliability Engineering (SRE) practices indicate that autoscaling alone might not be enough to make strong guarantees about a service's performance, specially under high load.⁵ In addition, Kubernetes does not yet support monitoring and management of a service's communication characteristics at the application layer (i.e., Layer 7), such as establishing per service rate limits and bandwidth quotas, which could be useful to help manage the performance and communication overhead of the video playback service.

Finally, the recent advent of the Function-as-a-Service (FaaS) and serverless computing cloud models,⁶ in which fined-grained services are deployed as serverless functions

5. <https://landing.google.com/sre/>

6. <https://martinfowler.com/articles/serverless.html/>

that are transparently managed and scaled by the cloud platform, and charged on a per usage basis, poses an additional challenge:

[C4] How to develop and manage self-adaptive microservice systems in hybrid deployment environments, composed of both serverful and serverless services.

For example, some of the video surveillance services, e.g., video processing, could be re-implemented and re-deployed as serverless services, using an FaaS cloud platform such as AWS Lambda.⁷ In that case, application developers would have not only to reconcile the adaptation requirements of both serverful and serverless services with the self-adaptation capabilities of their respective management platforms, but also to rethink their entire DevOps and business strategies, to account for the significant technical and economical differences between those two deployment models [10].

4.2 Control Loop Deployment

Control loops are crucial elements to realize the run-time adaptation of software systems. A typical self-adaptation control loop consists of four main activities, namely Monitor, Analyze, Plan, and Execute, all sharing a common Knowledge base, usually referred to as the MAPE-K reference model [1]. Control loops can be designed and deployed according to different control strategies, from a single centralized control component managing the whole system, to multiple control components managing different parts of the system and organized in a hierarchical or fully decentralized manner [2]. In the context of a microservice application, where each microservice is independently developed, deployed and managed at run time, selecting appropriate control strategies poses additional challenges.

An initial challenge here is:

[C5] How to determine the level of *distribution*, *visibility* and *granularity* necessary for deploying a microservice application's control components.

We see these three deployment dimensions as forming a control loop's *deployment space* for a microservice application. The distribution dimension concerns the physical allocation of control loop components to infrastructure resources. The visibility dimension, in turn, concerns whether the control loop components should be deployed at the application level, i.e., fully visible to developers, or at the infrastructure level, i.e., only partially visible to developers. Finally, the granularity dimension concerns whether control loop components should be deployed as a single monolithic service or decomposed into a collection of independently developed and managed microservices.

Developing and deploying self-adaptive microservice systems with those three dimensions in mind could help to establish fundamental tradeoffs with respect to multiple quality attributes. For instance, control loops for each of the video surveillance application services could be deployed in a fully decentralized fashion, which would increase their overall reliability and scalability. However, this

strategy would also make it harder to enforce application-wide adaptation constraints, as this would require each control loop to coordinate its actions with the other control loops, thus reducing their decision autonomy. Similarly, deploying control loops at the infrastructure level would help to promote a better separation between business and management services at run time, thus facilitating their reuse. However, this would also make them much harder to customize for specific adaptation needs, e.g., managing the expected accuracy of ML services, as most control loops provided at the infrastructure level support only a restricted set of adaptation models and mechanisms [11]. Finally, deploying control loop components as monolithic services would greatly simplify their packaging and management at run time. However, this strategy would also force all control components to share the same version and release rate, thus severely compromising their continuous improvement to satisfy evolving adaptation needs.

Another related challenge is:

[C6] How to reconcile the selected control loop deployment strategies with those supported by the underlying infrastructure management platform.

For instance, in Kubernetes self-healing and autoscaling controllers are deployed in a mostly centralized manner, as part of its master node. Therefore, if each microservice is to be managed by a fully independent control loop, there should be multiple instances of such types of controllers deployed, one for each microservice. In addition, those individual controllers might still have to coordinate their actions at the application level to resolve potential requirement conflicts, as discussed in Section 4.1, which Kubernetes currently does not support.

4.3 Continuous Delivery

We identify three main adaptation scenarios concerning the use of CD practices and tools in the context of a self-adaptive microservice system. The first scenario is the run-time adaptation of the CD pipelines themselves. In that regard, a key challenge is:

[C7] How to determine monitoring and adaptation mechanisms to dynamically adjust the transition events and conditions across the stages of each microservice CD pipeline.

For instance, to create a CD pipeline developers typically need to define the events and conditions under which each stage of the pipeline can be executed. This is usually done by manually expressing the trigger events and test requirements of each stage, so that the target application can be automatically progressed from one stage to the next. In this scenario, a self-adaptive pipeline would have to be able to monitor the application behavior at each stage, so as to dynamically adjust its trigger events and/or test requirements to cope with uncertainties during its execution, e.g., the application repeatedly failing to meet the requirements of one particular stage due to the necessary data or infrastructure resources being unavailable.

Another challenge related to this scenario is:

7. <https://aws.amazon.com/lambda/>

[C8] How to reconcile the adaptation needs of each microservice CD pipeline with the self-adaptation capabilities offered by the underlying infrastructure and CD management platforms.

For instance, Kubernetes offers a rollout feature that can be used to automatically deploy a new service version without having to take the current version down. This feature could be useful to implement a *canary release* pipeline for the face recognition service. A canary release pipeline is responsible for automatically deploying a new service version to production in parallel with the most recent stable version of that service, so as to gradually replace instances of the stable service version with instances of the new (canary) version based on a given set of testing criteria [9]. However, this solution would still require service developers to manually invoke Kubernetes's rollout feature every time there's a new service version to be released. Alternatively, developers could benefit from a full-fledged CD management tool, e.g., Spinnaker, to fully automate the execution of that canary release pipeline. But even in that case developers would still need to provide Spinnaker with the exact conditions and trigger events required to deploy the canary release into production.

The second scenario is the adaptation of the microservices' own adaptation requirements. In this scenario, a key challenge is:

[C9] How to dynamically adjust the adaptation requirements of each microservice according to the needs of each CD stage.

For instance, to streamline testing of the autoscaling capabilities of the video processing service during staging, the CD platform could automatically adjust that service's performance requirements, so as to trigger its autoscaling features earlier and more often than during normal production.

Finally, the third scenario concerns treating the self-adaptation mechanisms of a self-adaptive microservice system as first-class DevOps entities. In that respect, a further challenge is:

[C10] How to establish an effective strategy for continuously delivering self-adaptation mechanisms in production in a self-adaptive microservice system.

For instance, each self-adaptation mechanism used by the video surveillance application, e.g., autoscaling and self-healing, could be independently tested and deployed in its own CD pipeline. Such a strategy would allow a more systematic reuse of self-adaptation models and mechanisms across business services, at the expense of having a more complex CD process.

4.4 Testing

Conducting extensive validation tests with microservices before every deployment is not feasible due to the high frequency of their releases. Instead, microservice quality assurance is often compensated, or even replaced by, fine-grained monitoring techniques in production environments

exposed to real workloads. In this way, failures can be monitored and quickly corrected by pushing new releases into production [8].

A key challenge in this context is:

[C11] How to determine testing models and mechanisms to assess the fundamental quality attributes of each microservice and of the overall application from a self-adaptation perspective.

For instance, developers of the video surveillance application may need to select different testing mechanisms to assess the self-adaptation features of different microservices, e.g., image benchmarking and load generation for testing the autoscaling features of the face recognition service, and fault injection for testing the self-healing features of the video playback service. In addition, the need for those testing mechanisms may vary across CD stages and pipelines, e.g., autoscaling tests of the face recognition service may be run only during staging, while more critical self-healing tests of the video playback service may be run all the way to production.

Another challenge related to testing is:

[C12] How to integrate the systematic testing of microservices, including their self-adaptive behaviors, within the context of existing CD practices.

For example, fault injection tests created to assess the self-healing features of the video playback service in production could be incorporated by the underlying CD platform, as another condition upon which to decide whether the latest version of that service is ready to be promoted from canary release to normal production.

5 NEW DIRECTIONS

We now suggest promising new directions to address most of the challenges identified in the previous section.

5.1 New Adaptation Mechanisms

Regarding challenges C1 and C3, a practical way of implementing novel self-adaptation solutions tailored for the context of microservice applications would be by extending the management API provided by current container orchestration tools, like Kubernetes. For instance, one could easily build on the rollout and rollback features provided by Kubernetes to develop a self-adaptive service fallback mechanism that could be customized to meet different adaptation requirements.

The basic idea is to create multiple fallback versions for each microservice (e.g., low fidelity version, high accuracy version) and pack them as separate Docker images. The self-adaptive fallback mechanism could then be configured to use Kubernetes to automatically update the current image of a given microservice to one of its fallback images under certain system or environment conditions (e.g., update the face recognition service to its high accuracy image whenever the system is underutilized, and rollback that service to its original image whenever the load reaches a certain threshold).

Similarly, one could build on the variety of communication-related monitoring and management features provided by so-called *service mesh* tools [3], to create novel connector-oriented self-adaptation mechanisms. For instance, for some security attacks, the traffic management features provided by a service mesh tool like Istio⁸ could be leveraged by a self-protection mechanism to automatically strengthen the encryption parameters of the communication protocols being used by the video surveillance application, without the need to change or restart any of its services.

Another interesting use of a service mesh for self-adaptation, and that could be useful to address challenges C11 and C12, would be to dynamically adjust the *retry* and *circuit-breaking* parameters [3] used by clients of an overloaded service, so as to reduce the service's incoming traffic and thus prevent cascading failures from propagating throughout the system. This feature might be particularly useful for mission-critical self-adaptive systems.

Finally, specifically regarding challenge C2, application developers may rely on recent microservice orchestration languages (e.g., Jolie [12]) as the locus for defining and enacting application-aware adaptations. However, this decision may be seen as counterintuitive with respect to some well-established microservice principles, such as the autonomy of each microservice team to choose the implementation technologies that best fit their needs and expertise [3].

5.2 New Control Loop Deployment Structures

From a control loop deployment perspective, as discussed in the context of challenges C5 and C6, a microservice system's control components should ideally be deployed in a fully decentralized fashion, with each microservice being managed by its own local controller. However, this solution would make it harder for the local controllers to monitor and manage application-wide quality attributes. Having a centralized controller dedicated to managing application-level quality concerns would be a more straightforward solution in that regard, but this would also have the downside of creating a single point of failure and ultimately could compromise the application's overall availability. In practice, microservice developers may choose from a variety of intermediate solutions between those two extremes, e.g., by logically grouping services according to their business and/or quality affinity, and then having those services being collectively managed by independent yet application-aware group controllers organized in a hierarchical or fully decentralized structure.

Another important issue is the decision about whether microservice developers should have any responsibility in developing and managing control components, as suggested above. Having control components explicit in the design and development of a self-adaptive microservice system may contribute to further increase the system's overall complexity. Another drawback is that this decision might compromise the delivery autonomy of individual services, as service developers would have to negotiate before every new release in case their code bases share the same control components. We advocate a middle ground solution, in

which all control components are developed, tested and released as part of the underlying infrastructure management platform, but are still exposed to service developers during the later phases of their services' CD pipeline [11].

5.3 New Continuous Delivery Strategies

The question of how to integrate self-adaptation capabilities with DevOps, discussed in the context of challenges C7-C9, deserves special attention from self-adaptive microservice system developers. Here we discuss two possible CD strategies and their trade-offs.

One first strategy is to create a new dedicated CD pipeline to deliver into production all self-adaptation components used by the application. Having a dedicated self-adaptation pipeline would have the advantage of running it in parallel with the other business pipelines, thus preserving their autonomy. On the other hand, this solution would prevent business developers from directly improving the self-adaptation components used by their services, as this would require negotiating with the developers responsible for the self-adaptation pipeline. Another drawback of this solution is that it would push critical integration tests to the end of each business pipeline, as business developers would not have immediate access to the latest version of their required self-adaptation components.

A second strategy is to have separate business and self-adaptation pipelines during the commit and build stages, but have the self-adaptation pipeline integrated with the other business pipelines as soon as they enter the testing stage. This would avoid delaying integration tests in the business pipelines. However, as a side effect, it could unnecessarily bloat all business pipelines, with business developers now being responsible for testing both business and self-adaptation components.

Those two strategies are but a small sample of the spectrum of CD alternatives developers of self-adaptive microservice systems can explore. In that direction, extending emerging Model-Driven Engineering (MDE) tools for microservices, e.g., AjiL,⁹ with CD support could be a key enabler to automate the integration of self-adaptation capabilities and DevOps.

5.4 New Testing Approaches

A number of recent testing approaches have been proposed specifically for microservices, by both industry, e.g., integration testing in production,¹⁰ and academia, e.g., trace-based error prediction and fault localization [13]. A natural new direction here, related to challenges C11 and C12, is to systematically integrate those emerging testing approaches within the context of a fully automated self-adaptive CD pipeline. In this way, the results of both online and offline tests could be incorporated as part of the set of dynamic events and conditions used by underlying CD platform to automatically test a given microservice release across multiple pipeline stages.

Another promising direction in this context is applying *chaos engineering* [14] principles to assess the resiliency of a

8. <https://istio.io/>

9. <https://github.com/SeelabFhdo/AjiL>

10. <https://labs.spotify.com/2018/01/11/testing-of-microservices/>

self-adaptive microservice system. In that regard, one could rely on such a tool as kube-monkey,¹¹ which can be used to randomly terminate service instances in a Kubernetes cluster, to develop a kind of *adversarial control loop* whose main goal is to disturb the self-adaptive behavior of a target microservice system. This adversarial control loop could then leverage existing self-adaptation models and techniques to monitor and respond to (i.e., counter act) the microservice system's adaptation actions, thus offering a powerful mechanism to test its resiliency in production [15].

6 CONCLUSION

We have looked into a number of practical challenges facing the development and management of self-adaptive microservice systems. We have also suggested potential ways to address most of those challenges, with a focus on current and emerging microservice practices and technologies. We hope our ideas contribute to promote a better understanding of the interplay between self-adaptive systems and microservices, thus providing a timely incentive for researchers, practitioners, and tool developers to tackle some of the issues raised here as well as others that might arise.

REFERENCES

- [1] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003.
- [2] D. Weyns, *Software Engineering of Self-Adaptive Systems: An Organised Tour and Future Challenges*. Springer, 2017.
- [3] P. Jamshidi *et al.*, "Microservices: The journey so far and challenges ahead," *IEEE Software*, vol. 35, no. 3, pp. 24–35, 2018.
- [4] A. Balalaie, A. Heydarnoori, and P. Jamshidi, "Microservices architecture enables DevOps: Migration to a cloud-native architecture," *IEEE Software*, vol. 33, no. 3, pp. 42–52, 2016.
- [5] S. Rajagopalan and H. Jamjoom, "AppBisect: Autonomous healing for microservice-based apps," in *7th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*, 2015.
- [6] A. R. Sampaio Jr *et al.*, "Improving microservice-based applications with runtime placement adaptation," *Journal of Internet Services and Applications*, vol. 10, no. 1, pp. 1–30, 2019.
- [7] A. Filieri *et al.*, "Control strategies for self-adaptive software systems," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 11, no. 4, p. 24, 2017.
- [8] R. Heinrich *et al.*, "Performance engineering for microservices: Research challenges and directions," in *Proc. of the 8th ACM/SPEC on Int. Conf. Perf. Eng. Companion*. ACM, 2017, pp. 223–226.
- [9] L. Bass, I. Weber, and L. Zhu, *DevOps: A Software Architect's Perspective*. Addison-Wesley Professional, 2015.
- [10] G. Adzic and R. Chatley, "Serverless computing: Economic and architectural impact," in *11th Joint Meeting on Foundations of Software Engineering (FSE)*. ACM, 2017, pp. 884–889.
- [11] N. C. Mendonça *et al.*, "Generality vs. reusability in architecture-based self-adaptation: The case for self-adaptive microservices," in *1st Int. Workshop on Architectural Knowledge for Self-Adaptive Systems (AKSAS)*, 2018.
- [12] C. Guidi *et al.*, "Microservices: A language-based approach," in *Present and Ulterior Software Engineering*. Springer, 2017, pp. 217–225.
- [13] X. Zhou *et al.*, "Latent error prediction and fault localization for microservice applications by learning from system trace logs," in *27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*. ACM, 2019, pp. 683–694.
- [14] A. Basiri *et al.*, "Chaos engineering," *IEEE Software*, vol. 33, no. 3, pp. 35–41, 2016.
- [15] J. Cámara *et al.*, "Robustness-driven resilience evaluation of self-adaptive software systems," *IEEE Trans. Dependable Sec. Comput.*, vol. 14, no. 1, pp. 50–64, 2017.

11. <https://github.com/asobti/kube-monkey>



Nabor C. Mendonça is a professor of applied informatics at the University of Fortaleza. From 2017 to 2018 he was a visiting scholar in the School of Computer Science at Carnegie Mellon University, Pittsburgh. His research interests include software engineering, distributed systems, self-adaptive systems, and cloud computing. Mendonça received a Ph.D. in computing from Imperial College London. Contact him at nabor@unifor.br.



Pooyan Jamshidi is an assistant professor at the University of South Carolina, Columbia. His research interests include software engineering, systems, and machine learning, with a focus on the areas of machine-learning systems. Jamshidi received a Ph.D. from Dublin City University, Ireland. Contact him at pjamshid@cse.sc.edu.



David Garlan is a professor and associate dean in the School of Computer Science at Carnegie Mellon University, Pittsburgh. His research interests include autonomous and self-adaptive systems, software architecture, formal methods, explainability, and cyberphysical systems. Garlan received a Ph.D. in computer science from Carnegie Mellon University. Contact him at garlan@cs.cmu.edu.



Claus Pahl is a professor of computer science at the Free University of Bozen-Bolzano, where he heads the Software and Systems Engineering Group. His research interests include software engineering in service and cloud computing, specifically migration, architecture specification, dynamic quality, performance engineering, and scalability. Pahl received a Ph.D. in computing from the University of Dortmund. Contact him at cpahl@unibz.it.