

Chameleons on Cloudlets: Elastic Edge Computing Through Microservice Variants

Julien Gedeon, Martin Wagner, Karolis Skaisgiris, Florian Brandherm, Max Mühlhäuser

Telecooperation Lab, Technische Universität Darmstadt, Germany

Email: {gedeon, wagner, skaisgiris, brandherm, max}@tk.tu-darmstadt.de

Abstract—Common deployment models for Edge Computing are based on (composable) microservices that are offloaded to cloudlets. Runtime adaptations—in response to varying load, QoS fulfillment, mobility, etc.—are typically based on coarse-grained and costly management operations such as resource re-allocation or migration. The services themselves, however, remain non-adaptive, worsening the already limited elasticity of Edge Computing compared to Cloud Computing. Edge Computing applications often have stringent requirements on the execution time but are flexible regarding the quality of a computation. The potential benefits of exploiting this trade-off remain untapped.

This paper introduces the concept of *adaptable microservices* that provide alternative variants of specific functionalities. We define service variants that differ w.r.t. the *internal* functioning of the service, manifested in different algorithms, parameters, and auxiliary data they use. Such variants allow fine-grained trade-offs between the QoS (e.g., a maximum tolerable execution time) and the quality of the computation. We integrate adaptable microservices into an Edge Computing framework, show the practical impact of service variants, and present a strategy for switching variants at runtime.

Index Terms—edge computing, microservices, computation offloading, approximate computing, service adaptation

I. INTRODUCTION

Edge Computing [1]–[3] is transforming our computing landscape towards enabling low-latency computations at locations proximate to users. Consequently, many platforms and frameworks exist that propose Edge Computing runtimes, e.g., built on lightweight container deployments [4]–[7]. In addition, many Edge Computing deployments follow the paradigm of microservices [8], [9]. This abstraction has many advantages, such as higher agility and flexibility in the development of individual services. At runtime, multiple microservices can be combined to form processing chains and, according to demands, individual services can be scaled in and out.

While existing works are able to adapt the *management* of the services, e.g., through the placement [10], [11], or migration [12], [13] of services, the services *themselves* and their *internal* functioning remain non-adaptable. More specifically, edge services are implemented to deliver functionality in one particular way and cannot vary *how* the functionality is provided, e.g., by providing different *variants* of a service. Those variants may, for instance, differ in the algorithms they use to perform a task. As another example, some services need additional auxiliary data, which can also be varied (e.g.,

by using different pre-trained models for machine learning applications).

Different variants of a microservice potentially have an impact on two metrics: (i) the computational complexity, reflected in the execution time and resource demand of a request, and (ii) the quality of result (QoR). The latter can be defined and measured in different ways, e.g., by the accuracy of the result, i.e., its deviation from a (numeric) optimum, or by the (subjective) perception of a user. Overall, these two metrics form a trade-off, in the sense that more accurate results typically require more computational effort, which leads to higher execution times and/or increased resource demands. On the other hand, if we are willing to sacrifice computational quality, we can perform the same tasks with fewer resources.

This observation is especially remarkable in the context of Edge Computing if we recall some of its characteristics. On the one hand, computing resources available in Edge Computing are less powerful compared to their cloud counterparts, making efficient computing an important requirement to cope with scarce resources. Similarly, achieving resource elasticity is more challenging in Edge Computing, since the total available resources at a given location are much more limited. On the other hand, many edge applications have stringent requirements on the overall latency. At the same time, such mission-critical applications can be flexible regarding the quality of the computation result. Examples can be found in the domain of image or video processing, and for recognition tasks. To illustrate the practical impact of inaccurate computations, Chippa et al. [14] surveyed different kinds of applications and found that, on average, applications spent 83 % of their execution time on computations that are error-tolerant. Users of AR/VR headsets, for instance, might be willing to accept a lower rendering quality if in turn swift rendering helps in minimizing motion sickness. Current Edge Computing frameworks, however, do not consider this trade-off between computation effort and the quality of the computation result, and hence, miss out on this optimization opportunity.

In this paper, we present the novel concept of *adaptable microservices* for Edge Computing. We re-define microservices as blueprints for the delivery of a particular functionality that can be adapted w.r.t. (i) the algorithms they use to perform a task, (ii) parameters, and (iii) auxiliary data required for the computation. The possible *variants* are implemented within the program code of a microservice and can be selected upon its instantiation. Additionally, through a dedicated control

channel, the current variant of a service instance can be changed at runtime. The selection of the specific variant can be made according to certain requirements, e.g., a maximum tolerable execution time or a minimum quality of result. Furthermore, by having service variants with varying resource requirements, service variants are a way to bring the much-valued resource elasticity of Cloud Computing to the domain of Edge Computing.

We argue that adapting service instances can be a way to avoid cold start latencies, as changes are applied to running services, instead of potentially re-starting a service or replacing it, e.g., to more powerful hardware that can speed up its execution. Service variations are applied to an individual microservice, but they also have to be considered in the context of a microservice chain. For example, changing a variant of one microservice might have a disproportionate impact on the overall quality or execution time of the entire service chain.

We propose including this concept of adaptable microservices in an Edge Computing framework. In such a system, clients submit an abstract definition of the desired microservice or service chain with their individual requirements regarding execution time and QoR to a controller, which in turn has to make the following decisions: (i) which service variant to choose for instantiation in each step of the chain, and (ii) the assignment of user requests to service instances (since multiple services in different variants might be available). Furthermore, the controller might choose to change the variant of a particular microservice at runtime, e.g., by switching the algorithm with which the service performs its task. Especially in cases where microservice instances are shared between multiple microservice chains, this becomes a non-trivial optimization problem, because users that share (parts of) a chain might have conflicting optimization goals.

In summary, the contributions of this paper are threefold:

- We propose the concept of *adaptable microservices* in an Edge Computing environment. We define the adaptability of microservices in three dimensions (algorithms, parameters, and auxiliary data).
- We present a concept for the integration of adaptable microservices into an Edge Computing execution framework, detailing several components required for the orchestration of those service variants across edge cloudlets and users.
- We demonstrate the practical impact of orchestrating adaptable microservices w.r.t. (i) profiling their execution time, (ii) the effects of different variables on the service, and (iii) how switching variants can adapt to changes in the request patterns.

The remainder of this paper is structured as follows. We review related work in Section II. Our approach for dynamic microservice adaptation is presented in Section III. Section IV details how we realize this concept in an Edge Computing framework. In Section V, we demonstrate the practical impact of microservice variants and their switching at runtime. Section VI gives an outlook on future work before we conclude in Section VII.

II. BACKGROUND & RELATED WORK

Our contribution explores microservice adaptations in the context of Edge Computing. This section provides background information and reviews related work about Edge Computing frameworks in general (Section II-A), the concept of microservices and their adaptations (Section II-B), and approximate computing (Section II-C).

A. Edge Computing Frameworks

Since the emergence of Edge Computing and the related concept of cloudlets [15], different implementations of this computing paradigm have been proposed [16]. For example, Carrega et al. [17] present a general middleware for Mobile Edge Computing, that considers resources in the network of telecom operators. The framework of Ferrer et al. [18] focuses more on ad-hoc resources in the vicinity of users.

Much of the research has been centered around management issues, such as the placement [10], [11], [19] of Edge Computing services. Furthermore, in dynamic Edge Computing environments, migrations of users and services need to be handled accordingly [12], [13], [20]. Some works like *NOMAD* [21] integrate both caching and service migration strategies in an Edge Computing platform.

Similar to our implementation (see Section IV-A), some other works [7], [22], [23] also employ a repository for offloadable parts that are then transferred to surrogates. Paradrop [7] is an Edge Computing platform that enables deploying third-party applications on WiFi access points. *CloudPath* [22] is restricted to stateless functions. Both Paradrop and CloudPath do not allow the chaining of services. Bhardwaj et al. [23] present *AirBox*, a platform based on backend-driven onloading of functions to cloudlets. Compared to our approach, none of these works allow adaptations to the *internal* functioning of the provided services. Wang et al. [24] propose workload reduction as one method to increase scalability and elasticity for wearable cognitive assistance applications. In comparison, our three dimensions of adaptation can be applied to a wider range of applications.

B. Microservices, Service Adaptations, and Service Variants

Microservices are a contrasting paradigm to monolithic software. Following the microservice paradigm, parts of an application are developed and deployed independently [25]. The benefits of microservices, e.g., regarding DevOps [26], [27] or scalability [28], [29] have been widely recognized. However, developing applications as microservices also brings new challenges and concerns [30], [31], e.g., with regards to an increased operational complexity and testing efforts. Dragoni et al. [32] provide a more in-depth introductory survey about the general concept of microservices. Although the granularity of a microservice is not clearly defined [33], [34], microservices are typically characterized as small parts of an application with limited responsibilities, often restricted to performing a single task.

Adaptation of services has been explored in the broader context of service-oriented architectures (SOA) and Web Services (WS) [35]. Chang et al. [36] present a survey of common adaptation methods in service-oriented computing. Hirschfeld and Kawamura [37] define adaptability in three dimensions: *what* (e.g., computation/behavior or communication), *when* (e.g., at compile time or run time), and *how* (e.g., composition or transformation).

From a software engineering point of view, variants of services can be realized using *Software Product Lines* (SPL). SPLs are a development approach for re-usable and interchangeable software [38]. SPLs are characterized by their variability [39] and this variability can for instance be represented with *feature models* [40]. Based on such feature models, Sanchez et al. [41] present a heuristic-based method for the selection of an optimal configuration. Dynamic software product lines are capable to adapt, e.g., to user requirements or resource constraints [42]. As an example, Weckesser et al. [43] examine the reconfiguration of dynamic software product lines. Reconfiguration is done based on consistency properties and learned performance-influence models. The authors however do not consider service chains, and hence, cannot capture the interdependencies of adapting multiple services in a service chain.

More recent works have proposed adaptations for microservices in the context of the IoT. Gholami et al. [44] propose the usage of different versions of a microservice (lightweight or heavyweight), primarily for the purpose of scaling the application. Kannan et al. [45] present *GrandSLAM*, a microservice execution framework aimed to maximize throughput and reduce SLA violations. They do not modify the microservices themselves but instead change the request distributions. It is worth noting that these techniques can be used in conjunction with our proposed approach. Mendonça et al. [46] discuss the trade-off between generality and reusability in self-adaptive microservices. Bhattacharya and De [47] survey adaptation techniques in computation offloading, considering only the degree of concurrency and workload heterogeneity as variations in the applications. Some works present adaptation models for specific applications, e.g. streaming analytics [48], or to realize fault tolerance [49]. Others adapt the granularity of the services and not the underlying functionalities [34]. In contrast, we present a general concept for the adaptation of the internal functioning of microservices.

C. Approximate Computing

Approximate computing trades computation quality with the required effort to perform that computation [50]. The motivation to use approximate computing stems from the fact that in many problem domains of science and engineering, exact results are not required, but only results that are *good enough*. Examples can be found in the domain of digital signal processing, multimedia, and data analytics. Besides algorithmic resilience, users are also tolerant of inaccurate results. Examples are search results in information retrieval or the quality of images and video streams. In addition, the

usage context might also influence the required computation quality [51].

At the top level, we can distinguish between hardware and software approaches for approximate computing [52]. Hardware approaches work by introducing imprecise logic components [53], [54] or using techniques like voltage overscaling [55]. In Edge Computing, we cannot implement approximate computing on a hardware level, given that we opportunistically leverage existing, heterogeneous devices over which we have no direct control.

One example of application-level approximate computing is *FoggyCache* [56]. The authors propose to reuse computation results across devices, based on the observation that similar contextual properties map to the same or similar outcome. Perez et al. [57] have examined the latency-accuracy trade-off in MapReduce jobs when applying approximate computing. Chippa et al. [14] conduct a study in which they analyze the resilience of different applications to result inaccuracies. As demonstrated in [58], different combinations of approximate computing techniques can be combined. Their results suggest that up to 50 % in execution time can be saved while producing acceptable results. Other works have demonstrated the potential impact of approximate computing in different application domains, e.g., iterative methods [59], image compression [60], artificial neural networks [61], and deep learning [62].

Few previous works exist that apply approximate computing to domains that are related to Edge Computing. Zamari et al. [63] combine approximate computing with Edge Computing in an IoT scenario where sensor data is to be sent to the cloud for analytics. In-transit edge nodes contribute to the analytics by carrying out intermediate computations. This is coupled with approximate computing techniques on a software level, such as reducing the number of iterations or skipping certain parameter values. Wen et al. [64] employ a similar approach. They present *ApproxIoT*, combining approximate computing (by using only samples of a raw data stream) with hierarchical processing. Schäfer et al. [65] introduce several metrics for the *quality of computation* (QoC), for example, speed, precision, reliability, costs, and energy. They extend their *Tasklet* system [66]—an offloading middleware for distributed computing—to provide execution guarantees w.r.t. these QoC metrics. Compared to our adaptations, they do so not by modifying the internal functioning of the computation unit but by controlling their distribution.

In a broader context, Pejovic et al. [67] outline the challenges for approximate computing on mobile devices with a focus on the users' needs. Similarly, Machidon et al. [51] have noted that the field of approximate computing for mobile devices still lags behind its counterparts in the desktop and server environment. Using the example of mobile video decoding, the authors demonstrate how the acceptable quality degradation can vary according to the user's current context.

III. THE CONCEPT OF ADAPTABLE MICROSERVICES

Our approach is based on the concept of microservices, i.e., software components that execute a specific task. Applications

are typically composed of multiple such services. As a special case, services can form *service chains*, in which the output of one microservice is the input for a subsequent microservice. In the context of Edge Computing, this development method is especially useful, as it allows for fine-grained offloading decisions.

We propose the dynamic *adaptation* of microservices at runtime. We use the term *service adaptation* to refer to the *internal functioning* of the microservices. This definition stems from the observation that a particular functionality can be implemented in different ways, leading to many possible *service variants* between which we can switch at runtime. Triggering a switch can be done via an interface that the service exposes, e.g., to a controller that is responsible for orchestrating the services. This adaptation is orthogonal to other runtime optimizations that can be made in order to provide certain guarantees, e.g., the scaling or migration of microservices to meet execution time guarantees in view of an increased system load. Migration strategies, however, introduce a considerable overhead, as program code (and possibly execution environments and state) have to be transferred. Compared to costly migration strategies, we argue that our approach is a suitable alternative because it allows for quick reconfiguration of instance variants and, therefore, service instances can be kept active for a longer period of time.

Contrary to previous approaches, e.g., in the domain of approximate computing (see Section II-C), our concept of adaptable microservices combines the following three characteristics: (i) we adapt the internal functioning of a microservice, i.e., we operate on the application level and adaptations are implemented in the program code of the microservices, (ii) we propose adaptations in three general dimensions, and (iii) we envision a control entity that automatically selects and changes the service variants at runtime.

We make microservices adaptable in the following three dimensions:

- 1) **Algorithms:** A task can typically be performed by a variety of algorithms. Those not only differ in their runtime complexity, and hence, result in varying execution time, hardware requirements, and energy consumption, but also in their suitability for different applications. Taking the example of compressing an image, some compression algorithms are better suited for photographs while others perform better on vector graphics.
- 2) **Parameters:** Parameters are variable inputs to the microservice that influence its execution behavior. We model parameters as key-value pairs. Parameters can, for example, customize the algorithm that is used. Taking the same example of image compression, the desired image quality would be a parameter for such a microservice. Parameters can also be used to explicitly limit the execution time of a microservice, e.g., via loop perforation¹ [68].

¹loop perforation refers to skipping certain iterations in a loop or breaking the loop after a number of iterations.

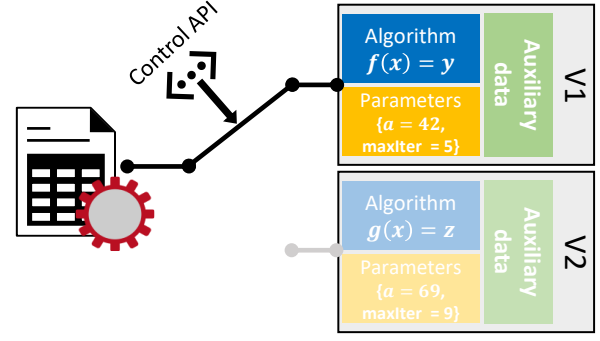


Figure 1. Switching between variants of an adaptable microservice

- 3) **Auxiliary Data:** Some algorithms require auxiliary data to function. This data is often retrieved from external sources. An example in the domain of machine learning are pre-trained models. This auxiliary data can also influence the execution time and the computation result. For example, in recognition tasks performed by neural networks, more complex models produce more accurate results, but require more computing resources or take longer to complete the task.

Figure 1 visualizes the concept of adaptable microservices. A given service might be variable in one or more of the aforementioned dimensions. We define the possible combinations of all three adaptation dimensions as *service variants*. As an example, the service in Figure 1 has two different variants.

Definition 1. Given a set of implemented algorithms \mathcal{A} , parameters \mathcal{P} , and auxiliary data \mathcal{D} for a microservice, a **service variant** Var_M of a microservice M is defined as $Var_M \subseteq \mathcal{A} \times \mathcal{P} \times \mathcal{D}$ with $p_i = v_i, i = 1 \dots n$ as values for the parameters. Note that \mathcal{A} and \mathcal{D} are finite sets, whereas \mathcal{P} typically is an uncountable set, e.g., in case the parameters contain real numbers.

We assume that there are no variants across a service chain that are mutually exclusive. Should one want to consider this case, constraint solvers can be used for selecting valid variants [69]. We further assume that each of the variants is implemented in the microservice. For example, if a microservice can be implemented using different algorithms, all those algorithms are included in the source code of the service. At any given time, a service maps its current variant to an internal state that determines how it is executed when requests are processed.

The different service variants impact the result of the computation in two ways. First, the execution time varies, e.g., when less complex algorithms are invoked or loop iterations are skipped. Naturally, this leads to a reduction in energy consumption of the cloudlet which executes the microservice. Second, service variants impact the quality of result (QoR). Depending on the application, QoR needs to be defined differently. We can divide QoR-metrics into two categories: (i) user-centered and (ii) numeric. For user-centered metrics,

techniques like questionnaires or focus groups can be used to assess the perceived quality of result. Note that this might not only vary from one user to another but also might depend on the usage context (as noted in [51]). As a numeric metric, we can for example quantify the error in the computation, i.e., the deviation from a numeric optimum or the accuracy of the result.

IV. REALIZATION OF THE CONCEPT AT THE EDGE

A. Integration into an Edge Computing Framework

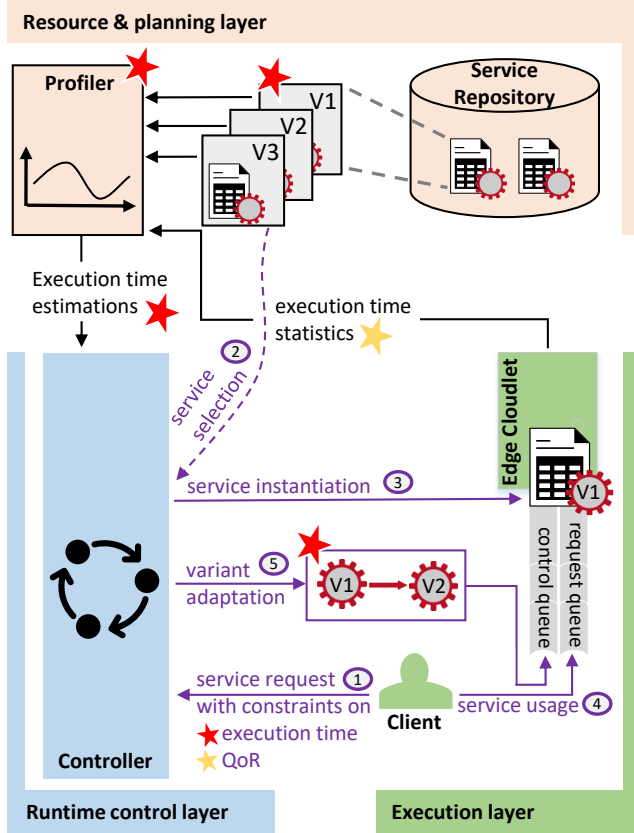


Figure 2. Adaptable microservices in an Edge Computing framework

To realize our concept, we implement the required mechanisms for adaptable microservices into our previously developed Edge Computing framework [70]. Figure 2 depicts an overview of this design. Components that are specifically related to the adaptable microservices are marked with a red star in Figure 2. Parts of the implementation that we leave open for future work are marked with a yellow star.

We structure the design of our system into three layers: (i) a resource and planning layer that provides the adaptable microservices and profiling of the services, (ii) the runtime control layer that manages the variants of the services, and (iii) the execution layer, where the service variants run on the edge agents. In our implementation, microservices are encapsulated as Docker containers that run on edge cloudlets. The microservices are made available through a repository. Users address requests for service execution to a *controller* that

is responsible for orchestrating the services. Communication with the microservices is done via asynchronous message queues. Our implementation uses *RabbitMQ* as a message broker.

Profiling and monitoring of adaptive services: Variants of a microservice are included in its implementation. For each variant, an offline profiler creates a model to estimate the execution time, given different input sizes and underlying hardware on which the service is executed. Section V-A will demonstrate the accuracy of such a profiling. This information serves as a basis for the controller for choosing suitable variants at the start of a microservice or changing the variants of running services. Given the heterogeneity of execution environments, not all possible hardware configurations can be considered. Hence, this information should be gradually updated with collected runtime statistics from the agents.

Control flow: Users can submit their requests for the execution of a service with a constraint on the execution time or the quality of result (shown as ① in Figure 2). Based on these constraints and the information from the profiler, a suitable service variant is selected ②, instantiated ③, and can then serve user requests ④. Note that for simplicity reasons, the figure only depicts a single microservice. For service chains, the decision-making process is made across all services in the chain. The question of how to place and spread the placement of individual microservices belonging to a service chain is beyond the scope of our paper. Previous works (e.g., [71]) have investigated this problem to great length.

Changing service variants: During the execution of a microservice, its variant can be changed. This is done by issuing requests to a dedicated *control queue* for each microservice instance. As an example, in Figure 2, the service variant is changed from V1 to V2 ⑤. Microservices implement a listener for incoming requests on the control queue and change their variant accordingly. This adaptation at runtime can be done for a number of reasons, e.g., when a constraint on the execution time cannot be met, a service might be instructed by the controller to switch to a variant that produces less accurate but faster results. Section V-C will explore the practical impact of variant switching.

B. Demo Microservices

We demonstrate our approach by using six individual microservices and two service chains.

1) *Individual microservices:* We implement the following adaptable microservices, summarized with their different variants in Table I:

Face detection: This microservice detects faces in a given picture. Its variants differ in algorithms and parameters. For the algorithms, we use two different types of cascade classifiers available in OpenCV: (i) *LBP* and (ii) *Haar*. In general, LBP is faster but produces less accurate results. The microservice furthermore expects two parameters: (i) *scale-factor* and (ii) *min-neighbors*. The first parameter determines the scaling between two levels of upscaling or downscaling (because both algorithms work only on predefined model dimensions).

Table I
SUMMARY OF SERVICE VARIANTS

Microservice	Algorithms	Variants Parameters	Auxiliary Data
Face detection	$\{LBP\text{-}Classifier, Haar\text{-}Classifier\}$	$\{scale\text{-}factor, min\text{-}neighbors\}$	\emptyset
Object detection	\emptyset	\emptyset	$\{faster_rcnn_inception_v2_coco, ssd_mobilenet_v1_coco, ssd_mobilenet_v1_fpn, ssd_mobilenet_v1_ppn, ssd_mobilenet_v2_coco, ssd_resnet50_v1_fpn, ssdlite_mobilenet_v2_coco\}$
Image compression	\emptyset	$\{compression\text{-}quality\}$	\emptyset
Image blurring	$\{Gaussian\ blur, Median\ blur\}$	$\{kernel\text{-}size\}$	\emptyset
Image upscaling	\emptyset	\emptyset	$\{psnr\text{-}large, psnr\text{-}small, noise\text{-}cancel, gans\}$
3D mesh reconstruction	\emptyset	\emptyset	$\{meshrcnn, pixel2mesh, sphereinit, voxelrcnn\}$

The second parameter *min-neighbors* specifies the minimum number of neighbors for candidate rectangles for those to be retained. Higher values for this parameter lead to fewer faces being detected but at the same time, this also decreases the number of false positives.

Object detection: This microservice uses *TensorFlow* to detect objects in a given image. The microservice uses different auxiliary data with pre-trained models². The models differ in their execution speed and mean average precision.

Image compression: Using the image encoding function of OpenCV, this microservice compresses a given input image using JPEG. As the only variation, the compression quality can be specified as a parameter.

Image blurring: Given an input image and an array of rectangular regions, this microservice blurs the given regions of the image. To perform the operation, we use OpenCV's blur function. The blurring can be performed by two different algorithms: (i) *Gaussian blur* and (ii) *median blur*. The Gaussian blur is a linear filter that is faster but does not preserve edges in the original image. In contrast, the median blur is a non-linear filter that is able to preserve edges. For both algorithms, a *kernel size* is used as a parameter to determine the size of the convolution matrix.

Image upscaling: This microservice produces an upscaled image of the input image. It also aims at enhancing the quality of the upscaled image by using Residual Dense Networks (RDN). We use an existing Keras³-based implementation⁴ as a basis for our microservice. We use four different pre-trained models that are variants of auxiliary data: *psnr-large*, *psnr-small*, *noise-cancel*, and *gans*. Except for the *gans* model (which quadruples the resolution), these models double the

original image resolution.

3D mesh reconstruction: This microservice aims at reconstructing a 3D mesh representation of an object in a (2D) picture. We use the published code⁵ of Gkioxari et al. [72] as the basis for our microservice. Four different models are used as auxiliary data. Some reconstruct only the shape of the object while others use voxels to achieve a more realistic representation of the object.

2) *Microservices chains:* From the individual microservices we construct two service chains.

Face anonymization: Given an image as input, this chain anonymizes faces by blurring them. First, the original image is compressed. Afterward, a face detection is performed, outputting detected faces as rectangular coordinates to the next service. As a final step, the image blurring microservice blurs the regions returned by the face detection microservice. An illustrative example of this service chain is shown in Figure 3(a).

3D mesh reconstruction of upscaled images: This microservice chain first performs an upscaling of an input image and then reconstructs a 3D mesh from the upscaled image. Figure 3(b) illustrates an example execution of this chain.

V. ORCHESTRATING MICROSERVICE VARIANTS

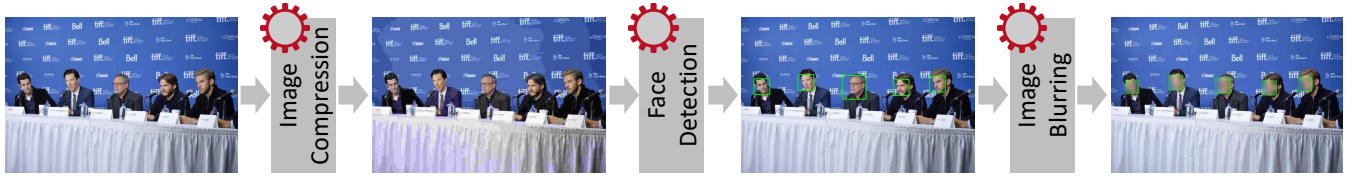
In this section, we study the practical impact of microservice variants. First, we study how accurately we can model the execution time of the microservices and which features are relevant for creating a model of the execution time that is as accurate as possible (Section V-A). We then study how different variables of service variants correlate and what their impact on the execution time and quality of result is (Section V-B). Lastly, we demonstrate how switching of service variants at runtime can help adapt to varying workloads in view of execution time constraints (Section V-C).

²https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md (accessed: 2020-04-22)

³<https://keras.io/> (accessed: 2020-04-21)

⁴<https://github.com/idealo/image-super-resolution> (accessed: 2020-04-16)

⁵<https://github.com/facebookresearch/meshrcnn> (accessed: 2020-04-16)



(a) Face anonymization



(b) 3D mesh reconstruction of upscaled images

Figure 3. Microservice chains

A. Execution Time Estimation

Table II
AWS INSTANCE TYPES USED FOR BENCHMARKING

Type	vCPUs	Clock rate	Memory
t2.micro	1	2.5 GHz ^a	1 GB
t2.small	1	2.5 GHz ^a	2 GB
t2.medium	2	2.3 GHz ^b	4 GB
t2.large	2	2.3 GHz ^b	8 GB
t2.xlarge	4	2.3 GHz ^b	16 GB
t2.2xlarge	8	2.3 GHz ^b	32 GB
c5.large	2	3.4 GHz ^c	4 GB
c5.xlarge	4	3.4 GHz ^c	8 GB
c5.4xlarge	16	3.4 GHz ^c	32 GB
r5d.large	2	3.1 GHz ^d	16 GB
r5d.2xlarge	8	3.1 GHz ^d	64 GB

^aIntel Xeon Family

^bIntel Broadwell E5-2695v4

^cIntel Xeon Platinum 8124M

^dIntel Xeon Platinum 8175

Methodology and experimental setup: To estimate the execution time of microservices, we build regression models using supervised learning methods. We use three variants of estimators implemented in *scikit-learn*⁶, a machine learning library for Python: (i) decision trees regressor, (ii) random forest regressor, and (iii) extra tree regressor. For each estimated model, the R^2 -score is computed to assess its quality. This metric gives an indication of how accurate the model is. To analyze the impact of different underlying hardware configurations, we run this evaluation on different AWS EC2 instance types as summarized in Table II. They differ in CPU and memory configuration and are optimized for either general-purpose (t-type instances), computing (c-type instances), or memory-intensive (r-type instances) applications. For each instance type, we run benchmarks of the face detection and object detection microservice. During the execution of the

microservices, we also record statistics on available hardware resources and system load. Those will serve as possible features to build a model of the execution time (for instance, to be able to predict the execution time given different load levels on the system). For face detection, we use the two different face detection algorithms, *LBP* and *Haar* [73]. The *scale-factor* parameter is varied from 1.0 to 1.9 in 0.1 increments and values for *min-neighbors* are varied from 1 to 10. We use 46 160 different images from the *WIDER FACE*⁷ dataset. For the object detection, we use two different models (*ssd_mobilenet_v1_coco* and *faster_rcnn_inception_v2_coco*) on the *val2017* dataset included in the *Coco Datasets*⁸. For each instance type, we list the combination of features and regression methods that lead to the highest R^2 -score. Features can be properties related to the variant of the microservice (e.g., a parameter) or attributes of the machine where it is executed. Table III shows the results for the face detection and Table IV the results for the object detection. In the tables, the features are ordered in decreasing order of importance, i.e., to what extent they contribute to the prediction of the execution time.

Impact of the machine types: From the results, we can observe that especially with the more powerful machines, we can achieve high R^2 -scores and, hence, a high accuracy of the model. For less powerful types of machines, e.g., *t2.micro* and *t2.small*, we get much lower scores. This is likely due to a greater variance in execution times that happens because *t2*-type instances are so-called *burstable instances*, i.e., if the system is overloaded, the CPU performance of the virtual machine is temporarily increased. Since this is likely to happen with the least powerful types we used, the high variance of execution times is due to the constant on-off switching of the performance boost.

Differences in estimators and features: As a second observation, in all but one case (face detection on a *c5.4xlarge*

⁶<https://scikit-learn.org> (accessed: 2020-04-13)

⁷<http://shuoyang1213.me/WIDERFACE/> (accessed: 2020-04-25)

⁸<http://cocodataset.org/> (accessed: 2020-04-25)

Table III
EXECUTION TIME ESTIMATION RESULTS FOR THE FACE DETECTION
MICROSERVICE

Instance	Features ^a	R^2 -Score
t2.micro	SF, DF-A, MN	0.4675
t2.small	CF, DF-A, MEM-A	0.4317
t2.medium	CF, CPU-U, DF-C	0.8717
t2.large	CPU-U, CF, SF, NF	0.9456
t2.xlarge	CPU-U, CF, SF, MN, NF	0.9842
t2.2xlarge	CF, SF, CPU-U, NF	0.9856
c5.large	CF, SF, CPU-U, NF	0.9887
c5.xlarge	CF, SF, CPU-U, NF	0.9914
c5.4xlarge	CF, CPU-U, SF, CPU-F, DF-C, MEM-A, MEM-U	0.9965
r5d.large	CF, SF, CPU-U, NF	0.9883
r5d.2xlarge	CF, SF, CPU-U, NF	0.9860

^aCF: classifier, CPU-F: CPU frequency, CPU-U: CPU usage,
DF-A: detected faces (absolute number),
DF-C: detected faces (correct percentage),
MEM-A: available memory, MEM-U: used memory,
MN: min-neighbors, NF: number of faces, SF: scale-factor

Table IV
EXECUTION TIME ESTIMATION RESULTS FOR THE OBJECT DETECTION
MICROSERVICE

Instance	Features ^a	R^2 -Score
t2.micro	n.a ^b	
t2.small	M, CPU-U, MEM-T	0.5651
t2.medium	MEM-AP, M, MEM-U, C, CPU-F	0.6938
t2.large	M, MEM-A, MEM-U	0.6827
t2.xlarge	M, MEM-A, MEM-T	0.6412
t2.2xlarge	M, MEM-AP, CPU-U	0.7690
c5.large	M, CPU-U, MEM-T	0.9970
c5.xlarge	M, CPU-U, MEM-T	0.9969
c5.4xlarge	M, CPU-U, MEM-T	0.9968
r5d.large	M, CPU-U, MEM-T	0.9968
r5d.2xlarge	M, CPU-U, MEM-T	0.9970

^aC: correctness, CPU-F: CPU frequency, CPU-U: CPU usage,
MEM-AP: available memory (percentage),
MEM-A: available memory (absolute), M: model,
MEM-T: total memory, MEM-U: used memory

^bhardware configuration not sufficient to run the microservice

instance), the *extra tree* regressor led to the highest R^2 -score. We can also observe great differences in the most relevant features for the execution time estimation. These differences can both be seen within one microservice, depending on the instance type, and across microservices. The estimation for the face detection mostly used the classification algorithm as the most relevant feature. With more powerful hardware, the classifier, the *scale-factor* parameter, and the current CPU usage are consistently ranked the most relevant features, while for less powerful machines, the *min-neighbors* parameter and the number of detected faces were included in the features.

Contrary to the face detection microservices, for the object detection, we can see a clearer division of relevant features depending on the instance type. While for *t2*-type instances,

the available memory is always a highly ranked feature (except for the *t2.small* instance), this changes in favor of the CPU utilization for *c*-type and *r*-type machines. Another difference is that the *t2*-type instances lead to significantly lower accuracies of the model, as shown by the R^2 -score.

Summary: In summary, this analysis of execution time estimators has shown that we are able to accurately profile the different variants of microservice. In an Edge Computing framework where adaptable microservices are integrated (see Section IV-A), this step would be performed offline and serve as base knowledge for runtime decisions. However, we could also observe that this estimation has to be tuned to the individual microservice w.r.t. the selection of the hardware and features that are used for the estimation.

B. Impact of Service Variants

Methodology: We measure the correlation between different variables that relate to the service variants and the outcomes of the computations. Most importantly, we want to assess the change in execution time. In addition, for the face detection algorithm and, consequently, for the face anonymization service chain, we also analyze the impact on the quality of the result.

To measure the pairwise correlation between variables, we use the *Kendall rank correlation coefficient* throughout this section. Contrary to other metrics for correlation, such as the *Pearson correlation coefficient*, it has the advantage that it does not assume a linear relationship between variables.

	face detection algorithm	min-neighbors	given faces	detected faces	correctness	compression quality	blurring algorithm	execution time
face detection algorithm	1.00	0.00	0.00	-0.10	-0.10	0.00	0.00	-0.68
min-neighbors		1.00	0.00	-0.16	-0.26	0.00	0.00	-0.08
given faces			1.00	0.65	0.27	0.00	0.00	0.09
detected faces				1.00	0.70	0.02	0.00	0.16
correctness					1.00	0.04	0.00	0.16
compression quality						1.00	0.00	0.05
blurring algorithm							1.00	0.05
execution time								1.00

Figure 4. Face blurring chain: correlation matrix of variants

Face anonymization service chain: For the first chain, we vary the image compression quality from 1–99 (in steps of 1). We use the two face detection algorithms as described before. The *scale-factor* parameter is set to a constant 1.2, and *min-neighbors* are varied from 0–9 (step size 1). For the final step, the blurring microservice, we use *gaussian blur* and *median blur* algorithms with a fixed *kernel size* of (23, 23). We select 21 images and manually label the correct positions of the faces. Hence, with a small degree of tolerance, besides the absolute number of detected faces, we can also compute a *correctness* value that serves as a metric for the QoR. For

each image and combination, we executed the chain five times and averaged the results.

Figure 4 shows the correlation matrix of the entire chain. We can observe that the highest correlation value is attained among the face detection algorithm and the execution time. To map this correlation to concrete numbers, on average, the execution time using the Haar classifier was 0.13 s, while for the LBP classifier it averaged to 0.08 s. This means that by changing the variant of the algorithm, we could achieve a reduction in the execution time of 38.46 %. However, this reduction in execution time comes at the cost of a reduced correctness value, which drops from 0.67 to 0.57 on average (-14.92 %). This provides a good example of the trade-off between the computation complexity (represented by the execution time) and the quality of result (represented by the correct recognition of faces).

Compared to the face detection algorithm, other variables related to the variants, i.e., *min-neighbors*, *compression quality*, and *blurring algorithm* correlate with the execution time with values of -0.08, 0.05, and 0.05, respectively. It is worth noticing that *min-neighbors* has a much more significant impact on the correctness (with a correlation value of -0.26) than on the execution time.

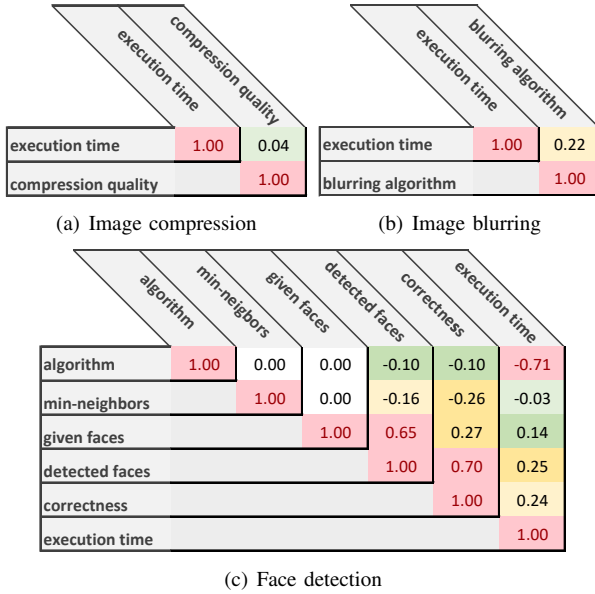


Figure 5. Correlation matrices for the individual service variants of the face anonymization chain

We also provide the correlation matrices of the individual services of this chain in Figure 5. Comparing Figure 5 with Figure 4 demonstrates the difference in correlation of a single microservice versus when this microservice is integrated into a chain. As an example, when executed alone, the blurring algorithm has a correlation value of 0.22 with the execution time but in the entire chain, this value drops to 0.05. A similar change in the correlation score can be observed for the face detection algorithm (-0.68 to -0.71).

3D mesh reconstruction of upscaled images: For both the image upscaling and 3D mesh reconstruction microservice, we use the four different variants of auxiliary data as previously described. As input data, we use 5 images from a dataset depicting furniture⁹. Because the mesh reconstruction microservice offers GPU support, we execute this service chain on an AWS EC2 *p2.xlarge* instance (Xeon E5-2686 v4, 61 GB RAM, Nvidia K80 GPU).

Figure 6 shows the correlation matrix for the entire chain and Figure 7 the matrices for the individual microservices. Note that for this microservice chain, we leave the exploration of suitable QoR-metrics for future work and focus on the execution times.

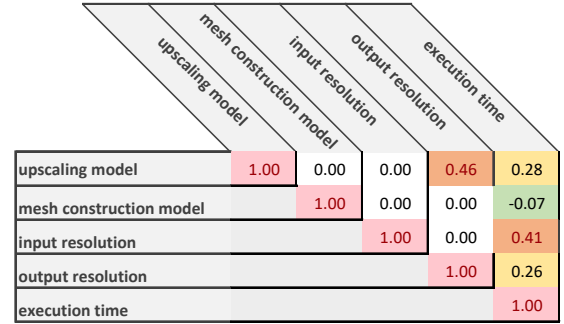


Figure 6. Mesh reconstruction chain: correlation matrix of variants



Figure 7. Correlation matrices for the individual service variants of the mesh reconstruction chain

The results show that the variants of the upscaling model have more influence than the different mesh reconstruction models (correlation scores of 0.28 and -0.07). As an example, the *psnr-small* model for image upscaling has an average

⁹<https://www.kaggle.com/akkithetechie/furniture-detector/data> (accessed: 2021-01-07)

execution time of 11.55 s while the *psnr-large* model averages to 58.94 s. The mean values for the *noise-cancel* and *gans* models are 63.93 s and 36.36 s, respectively. This means that by selecting another variant of an image upscaling model, we can reduce the execution time by up to 81.93 %. In comparison, the differences for the average execution times of the mesh construction models are smaller (41.26 s for *meshrcnn*, 42.01 s for *pixel2mesh*, 43.12 s for *sphereint*, and 44.38 s for *voxelrcnn*). Hence, here the maximum difference in execution time only amounts to 7.03 %. Naturally, there is also a strong correlation (0.41 and 0.26) of the execution time with the input and output resolution of the upscaled images.

C. Switching of Service Variants

Queue model and variant switching strategy: We model the processing of requests by a microservice as an *M/D/1 queue*. With this model, we are able to estimate the number of queued messages at which a switch to a faster microservice variant should happen, given a time constraint C that should not be violated. We assume that requests arrive according to a Poisson process with an arrival rate λ . Consequently, the inter-arrival times of requests follow an exponential distribution. We assume a deterministic service time D in our model, which is the average execution time of a microservice variant on a given hardware setup. Accordingly, the service rate μ can be calculated with the formula $\mu = \frac{1}{D}$.

Our strategy for variant switching tries to select the microservice with the highest execution time, so that a given time constraint C is not violated. This is based on the assumption that more complex microservices deliver better results. Furthermore, we assume that the length L of the request queue is known at any given time. The waiting time ω of a request is the sum of the time it has to wait in the queue ω_Q and the service time D of the microservice $\omega = \omega_Q + D, \omega_Q = L * D$. ω and C are the basis for the estimation of a threshold at which a microservice will switch to a less computing-intensive, and therefore, faster variant. To avoid a violation of C , the queue length L should not exceed a certain threshold T , so that the condition $\omega < C$ holds. T is calculated with C and D as $T = \lfloor \frac{C}{D} - 1 \rfloor$. To allow for fine adjustments of T , we introduce a dampening factor $T_{dampened} = \alpha T, \alpha \in (0, 1]$. A variant switch will be performed if $L > T_{dampened}$. $T_{dampened}$ is chosen in such a way that violations of C are minimized. Our variant switching strategy tries to avoid the request queue becoming unstable. The queue utilization ρ is used to determine the stability of a queue and is given as $\rho = \frac{\lambda}{\mu}$. If $\rho > 1$, i.e., if $\lambda > \mu$, then the queue becomes unstable because more requests arrive than the microservice can handle and the request queue will fill up. This will lead to violations of the execution time constraint as more requests in the queue lead to longer waiting times. To avoid this, our variant switching strategy tries to select a variant that can handle the expected request load, i.e., a variant with a sufficiently high μ so that $\lambda < \mu$. Furthermore, the selected variant should not lead to violations of C , i.e., for the average wait time of requests ω_{avg} for the selected variant

the condition $\omega_{avg} < C$ should hold. ω_{avg} can be calculated as follows:

$$\omega_{avg} = \frac{1}{\mu} + \frac{\rho}{2\mu(1-\rho)} = D + \frac{D\rho}{2(1-\rho)} \quad (1)$$

Methodology and experimental setup: For our experiments, we send a number of requests via a request generator to the request queue of a running microservice. The times between requests are drawn from an exponential distribution with parameter λ . We use two of our microservices described in Section IV-B for our experiments, namely face detection and image upscaling. We evaluate our variant switching strategy with different λ and measure the execution time of each request and the length of the request queue over the course of the experiments. We compare our results with the baseline queue lengths and execution times of microservice executions where the variant switching is disabled. The microservices are executed on a Lenovo ThinkCentre M920X Tiny (Intel Core i7-8700, 16 GB RAM, Ubuntu 18.04). Such small scale, energy-efficient, yet powerful devices are representative examples of possible edge resources, e.g., when publicly deployed in urban environments [74]

Experimental results: Figures 8(a), 8(b), 8(e) and 8(f) show the results of our variant switching experiments with the face detection microservice. In our experiments, the face detection microservice changes its algorithm if a variant switch is performed. The microservice starts with the *Haar-Classifer* and eventually switches to the *LBP-Classifer* as an algorithm for the face detection. We conduct two variants of the experiment for the face detection microservice. In each, we send 500 requests, define an execution time constraint of 500 ms, and set the dampening parameter to $\alpha = 1$. This results in a variant switching threshold of 6 requests in the queue. For the first experiment, we set the arrival rate to $\lambda = 15$, while for the second experiment, we increase the rate to $\lambda = 17$, so that the inter-arrival times of requests are shorter than in the first experiment. As a baseline for comparison, we also report on the results when there is no switching of microservice variants. In the experiment with $\lambda = 17$, the stability check in our switching strategy is disabled. Otherwise, the variant that uses the *Haar-Classifer* would not be chosen at all and a variant switch would not be performed.

Figures 8(a) and 8(e) show that as the execution time reaches the constraint and the queue length reaches the threshold, a variant switch to the *LBP-Classifer* is initiated at request number 77. The continued execution with the faster variant avoids a violation of the user-given execution time constraint of 500 ms. Without variant switching, the constraint is violated repeatedly. Similar results can be observed in Figures 8(b) and 8(f). Because of a greater λ compared to the first experiment, the execution time and queue length grow faster. This leads to a variant switch at request number 82. Without variant switching, the microservice will be overwhelmed with requests, which leads to a steady increase in both the execution time and queue length. As a consequence, the execution time

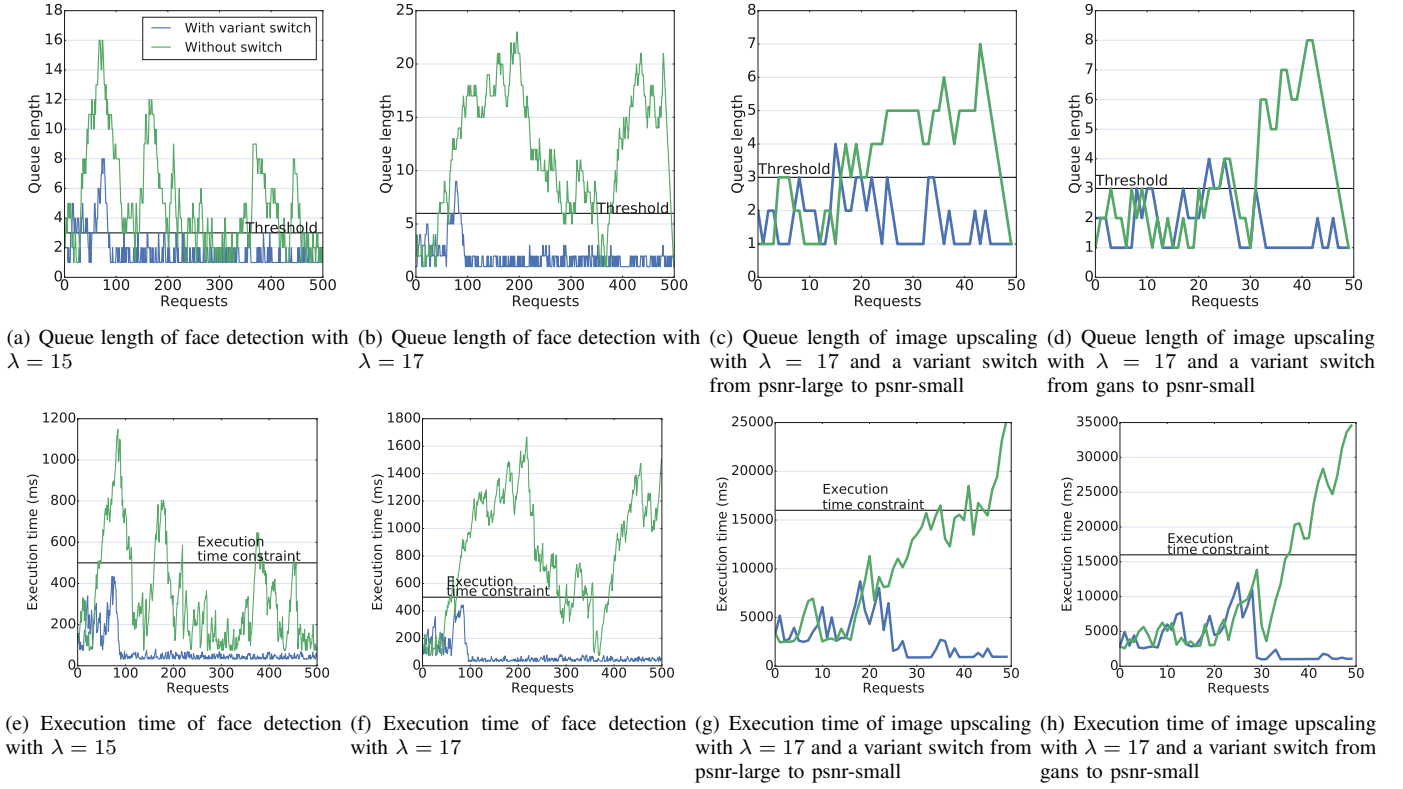


Figure 8. Queue length and execution time of face detection and image upscaling service variants

constraint is violated for most of the requests after the first violation.

Figures 8(c), 8(d), 8(g) and 8(h) show the results of our variant switching experiments with the image upscaling microservice. In these experiments, the microservice switches its variant by changing the pre-trained model that it uses. For our experiments, we used the models *psnr-small*, *psnr-large*, and *gans*. Of these three models, *psnr-small* leads to the shortest execution time while *gans* leads to the longest. We conducted our experiments with the image upscaling microservice two times. In each, we send 50 requests, define an execution time constraint of 16000ms, and set the dampening parameter to $\alpha = 1$. For the models *gans* and *psnr-large*, this leads to a threshold for the variant switching of 3. We use values of $\lambda = 17$ for both experiments. Again, we also include a baseline where no switch is performed for comparison.

Figures 8(c) and 8(g) show the experimental results of the image upscaling microservice. Because of the request arrival rate of $\lambda = 17$, the execution starts with the model *psnr-large* instead of *gans* and later switches to the model *psnr-small*. This decision is made because of the queue stability condition, so that an unstable request queue, where requests accumulate, is avoided. Without the variant switch, the constraint would be violated for several requests. Figures 8(d) and 8(h) show the results of our second experiment with the image upscaling microservice and $\lambda = 17$. For this experiment, we disable the aforementioned queue stability condition, so that we can enforce the execution of the microservice variant with the

gans model, leading to an unstable request queue. At request number 29 a variant switch is performed to the variant with the *psnr-small* model. This model has been chosen according to our condition, that the average waiting time (Equation (1)) for requests has to be lower than the execution time constraint. After the switch, constraint violations are avoided. Without the switch, the request queue would fill up which leads to constant violations of the execution time constraint.

Additionally, we compare our variant switching approach to an alternative approach, where a service would be restarted in the variant it is supposed to switch to. The process of restarting takes about 1.8s for the face detection service and 4s for the image upscaling service. The variant switch approach we use in our experiments takes about 17ms for both the face detection and image upscaling microservices.

Summary: We have shown that switching microservice variants at runtime can help ensuring execution time constraints. We demonstrated this with two microservices and a threshold-based switching strategy. It automatically decides when to switch the microservice variant and also decides which variant to switch to. This approach avoids expensive restarts of services in Edge Computing environments. Note that we plan to extend our switching strategy, so that switches to slower variants that deliver better results are performed when the queue has been stable enough for a sufficient amount of time. This way, execution time constraints can be met while also maximizing the QoR.

VI. DISCUSSION & FUTURE WORK

To demonstrate the benefits of adaptable microservices and present our switching strategy (see the previous Section V), we implemented a subset of the overall design shown in Figure 2. However, a number of practical issues remain for the global orchestration of service variants. Our work opens up the following questions for future work:

Hierarchical monitoring and control. To ensure that application-specific constraints w.r.t. the execution time and result quality are met, the execution of service chains needs to be monitored. Given the highly distributed nature of edge cloudlets, having only one centralized controller does not meet the scalability requirements of Edge Computing. Hence, we envision hierarchical monitoring and control mechanisms.

Network control layer. In a distributed Edge Computing system, not only the resources on the edge nodes and the microservices' complexity influence the execution time but also the network conditions and types of connections between the nodes (e.g., when the microservices of one service chain run on different nodes). Future work should take this into consideration in two aspects: First, fine-grained monitoring of network conditions can help in making runtime decisions for the placement and assignment of microservices. Second, we can extend the control itself to the network layer, e.g., by reserving bandwidth on links or using SDN to control the data flow between edge nodes.

Defining and weighting multiple QoR metrics. As we have noted, the quality of a computation can be defined in different ways. However, the interplay between user-perceived QoR and mathematical metrics for QoR is not well understood yet. It also remains unclear how both types of QoR should be weighted if they are part of one service chain. Furthermore, if a microservice instance is part of multiple service chains, this could lead to conflicts w.r.t. the individual optimization targets.

Defining service variants through SPLs. Software product lines allow for a more general modeling of application variants. Using this technique would also make it possible to model more complex dependencies between variants (e.g., when certain combinations of variants are mutually exclusive).

VII. CONCLUSION

Based on three properties of Edge Computing and its applications—constrained resources, tight constraints on the execution time, and flexibility regarding the quality of the computations—this paper proposed the general concept of *adaptable microservices*. Specifically, we defined microservices to be adaptable in three aspects, related to the *internal functioning* of the microservices. We designed the concept for the integration of adaptable microservices into an Edge Computing framework.

After having shown that we can accurately profile their execution times, we demonstrated the practical impact of adaptable microservice variants in relevant application domains of computer vision and image processing. Adaptable microservices allow trading the quality of computations for

lower resource utilization (manifested for example in a reduced execution time). We furthermore demonstrated how switching service variants at runtime can help adapting to changing request patterns. The proposed concept of microservice variants can help in mitigating the limited elasticity of Edge Computing by *adapting the services to the limitations of the execution infrastructure* and not vice versa.

ACKNOWLEDGEMENT

This work has been cofunded by the German Research Foundation (DFG) and the National Nature Science Foundation of China (NSFC) joint project under Grant No. 392046569 (DFG) and No. 61761136014 (NSFC), and as part of the Collaborative Research Center 1053 - MAKI (DFG). This work was supported by the AWS *Cloud Credits for Research* program.

REFERENCES

- [1] J. Gedeon, F. Brandherm, R. Egert, T. Grube, and M. Mühlhäuser, "What the Fog? Edge Computing Revisited: Promises, Applications and Future Challenges," *IEEE Access*, vol. 7, pp. 152 847–152 878, 2019.
- [2] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.
- [3] M. Satyanarayanan, "The Emergence of Edge Computing," *IEEE Computer*, vol. 50, no. 1, pp. 30–39, 2017.
- [4] C. Pahl and B. Lee, "Containers and Clusters for Edge Cloud Architectures - A Technology Review," in *Proc. of the 3rd International Conference on Future Internet of Things and Cloud*, 2015, pp. 379–386.
- [5] R. Morabito, V. Cozzolino, A. Y. Ding, N. Beijar, and J. Ott, "Consolidate IoT Edge Computing with Lightweight Virtualization," *IEEE Network*, vol. 32, no. 1, pp. 102–111, 2018.
- [6] K. Kaur, T. Dhand, N. Kumar, and S. Zeadally, "Container-as-a-Service at the Edge: Trade-off between Energy Efficiency and Service Availability at Fog Nano Data Centers," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 48–56, 2017.
- [7] P. Liu, D. Willis, and S. Banerjee, "ParaDrop: Enabling Lightweight Multi-tenancy at the Network's Extreme Edge," in *Proc. of the IEEE/ACM Symposium on Edge Computing (SEC)*, 2016, pp. 1–13.
- [8] I. Filip, F. Pop, C. Serbanescu, and C. Choi, "Microservices Scheduling Model Over Heterogeneous Cloud-Edge Environments As Support for IoT Applications," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2672–2681, 2018.
- [9] M. Alam, J. Rufino, J. Ferreira, S. H. Ahmed, N. Shah, and Y. Chen, "Orchestration of Microservices for IoT Using Docker and Edge Computing," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 118–123, 2018.
- [10] F. A. Salaht, F. Desprez, and A. Lebre, "An Overview of Service Placement Problem in Fog and Edge Computing," *ACM Computing Surveys*, vol. 53, no. 3, Jun. 2020.
- [11] T. Ouyang, Z. Zhou, and X. Chen, "Follow Me at the Edge: Mobility-Aware Dynamic Service Placement for Mobile Edge Computing," *IEEE JSAC*, vol. 36, no. 10, pp. 2333–2345, 2018.
- [12] S. Wang, J. Xu, N. Zhang, and Y. Liu, "A survey on service migration in mobile edge computing," *IEEE Access*, vol. 6, pp. 23 511–23 528, 2018.
- [13] L. Ma, S. Yi, and Q. Li, "Efficient service handoff across edge servers via docker container migration," in *Proc. of the IEEE/ACM Symposium on Edge Computing (SEC)*, J. Zhang, M. Chiang, and B. M. Maggs, Eds., 2017, pp. 11:1–11:13.
- [14] V. K. Chippa, S. T. Chakradhar, K. Roy, and A. Raghunathan, "Analysis and characterization of inherent application resilience for approximate computing," in *Proc. of the 50th Annual Design Automation Conference (DAC)*, 2013, pp. 113:1–113:9.
- [15] M. Satyanarayanan, P. Bahl, R. Cáceres, and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14–23, 2009.
- [16] K. Dolui and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," in *Proc. of the Global Internet of Things Summit (GIoTS)*, 2017, pp. 1–6.

- [17] A. Carrega, M. Repetto, P. Gouvas, and A. Zafeiropoulos, "A Middleware for Mobile Edge Computing," *IEEE Cloud Computing*, vol. 4, no. 4, pp. 26–37, 2017.
- [18] A. J. Ferrer, J. M. Marquès, and J. Jorba, "Ad-Hoc Edge Cloud: A Framework for Dynamic Creation of Edge Computing Infrastructures," in *Proc. of the 28th International Conference on Computer Communication and Networks (ICCCN)*, 2019, pp. 1–7.
- [19] S. Wang, M. Zafer, and K. K. Leung, "Online Placement of Multi-Component Applications in Edge Computing Environments," *IEEE Access*, vol. 5, pp. 2514–2533, 2017.
- [20] M. Chen, W. Li, G. Fortino, Y. Hao, L. Hu, and I. Humar, "A dynamic service migration mechanism in edge cognitive computing," *ACM Trans. Internet Techn.*, vol. 19, no. 2, pp. 30:1–30:15, 2019.
- [21] A. Pamboris, M. Baguena, A. L. Wolf, P. Manzoni, and P. R. Pietzuch, "Demo: NOMAD: An Edge Cloud Platform for Hyper-Responsive Mobile Apps," in *Proc. of the 13th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2015, p. 459.
- [22] S. H. Mortazavi, M. Salehe, C. S. Gomes, C. Phillips, and E. de Lara, "Cloudpath: a multi-tier cloud computing framework," in *Proc. of the ACM/IEEE Symposium on Edge Computing (SEC)*, 2017, pp. 20:1–20:13.
- [23] K. Bhardwaj, M. Shih, P. Agarwal, A. Gavrilovska, T. Kim, and K. Schwan, "Fast, Scalable and Secure Onloading of Edge Functions Using AirBox," in *Proc. of the IEEE/ACM Symposium on Edge Computing (SEC)*, 2016, pp. 14–27.
- [24] J. Wang, Z. Feng, S. A. George, R. Iyengar, P. Pillai, and M. Satyanarayanan, "Towards scalable edge-native applications," in *Proc. of the ACM/IEEE Symposium on Edge Computing (SEC)*, 2019, pp. 152–165.
- [25] M. Fowler, "Microservices - a definition of this new architectural term," <https://martinfowler.com/articles/microservices.html>, 2014, accessed: 2021-01-07.
- [26] 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.
- [27] H. Kang, M. Le, and S. Tao, "Container and Microservice Driven Design for Cloud Infrastructure DevOps," in *Proc. of the 2016 IEEE International Conference on Cloud Engineering (IC2E)*, 2016, pp. 202–211.
- [28] P. Jamshidi, C. Pahl, N. C. Mendonça, J. Lewis, and S. Tilkov, "Microservices: The Journey So Far and Challenges Ahead," *IEEE Software*, vol. 35, no. 3, pp. 24–35, 2018.
- [29] N. Dragoni, I. Lanese, S. T. Larsen, M. Mazzara, R. Mustafin, and L. Safina, "Microservices: How To Make Your Application Scale," in *Proc. of the 11th International Andrei P. Ershov Informatics Conference (PSI)*, 2018, pp. 95–104.
- [30] D. Taibi, V. Lenarduzzi, and C. Pahl, "Processes, Motivations, and Issues for Migrating to Microservices Architectures: An Empirical Investigation," *IEEE Cloud Computing*, vol. 4, no. 5, pp. 22–32, 2017.
- [31] J. Soldani, D. A. Tamburri, and W. van den Heuvel, "The pains and gains of microservices: A Systematic grey literature review," *Journal of Systems and Software*, vol. 146, pp. 215–232, 2018.
- [32] N. Dragoni, S. Giallorenzo, A. Lluh-Lafuente, M. Mazzara, F. Montesi, R. Mustafin, and L. Safina, "Microservices: Yesterday, Today, and Tomorrow," in *Present and Ulterior Software Engineering*, 2017, pp. 195–216.
- [33] S. Hassan, N. Ali, and R. Bahsoon, "Microservice Ambients: An Architectural Meta-Modelling Approach for Microservice Granularity," in *Proc. of the 2017 IEEE International Conference on Software Architecture (ICSA)*, 2017, pp. 1–10.
- [34] S. Hassan and R. Bahsoon, "Microservices and Their Design Trade-Offs: A Self-Adaptive Roadmap," in *Proc. of the IEEE International Conference on Services Computing (SCC)*, 2016, pp. 813–818.
- [35] M. P. Papazoglou, "Service-Oriented Computing: Concepts, Characteristics and Directions," in *Proc. of the Fourth International Conference on Web Information Systems Engineering*, ser. WISE '03, 2003, pp. 3–12.
- [36] S. H. Chang, H. J. La, and S. D. Kim, "A Comprehensive Approach to Service Adaptation," in *Proc. of the IEEE International Conference on Service-Oriented Computing and Applications (SOCA '07)*, 2007, pp. 191–198.
- [37] R. Hirschfeld and K. Kawamura, "Dynamic service adaptation," *Software Practice and Experience*, vol. 36, no. 11-12, pp. 1115–1131, 2006.
- [38] J. D. McGregor, L. M. Northrop, S. Jarrad, and K. Pohl, "Guest Editors' Introduction: Initiating Software Product Lines," *IEEE Software*, vol. 19, no. 4, pp. 24–27, 2002.
- [39] J. van Gorp, J. Bosch, and M. Svahnberg, "On the notion of variability in software product lines," in *Proc. of the Working IEEE/IFIP Conference on Software Architecture*, 2001, pp. 45–54.
- [40] D. Beuche and M. Dalgarno, "Software product line engineering with feature models," *Overload Journal*, vol. 78, pp. 5–8, 2007.
- [41] L. E. Sanchez, S. Moisan, and J. Rigault, "Metrics on feature models to optimize configuration adaptation at run time," in *Proc. of the 1st International Workshop on Combining Modelling and Search-Based Software Engineering (CMSBSE)*, 2013, pp. 39–44.
- [42] S. Hallsteinsen, M. Hinchey, S. Park, and K. Schmid, "Dynamic Software Product Lines," *Computer*, vol. 41, no. 4, pp. 93–95, 2008.
- [43] M. Weckesser, R. Kluge, M. Pfannemüller, M. Matthé, A. Schürr, and C. Becker, "Optimal reconfiguration of dynamic software product lines based on performance-influence models," in *Proc. of the 22nd International Systems and Software Product Line Conference (SPLC)*, 2018, pp. 98–109.
- [44] S. Gholami, A. Goli, C.-P. Bezemer, and H. Khazaei, "A Framework for Satisfying the Performance Requirements of Containerized Software Systems Through Multi-Versioning," in *Proc. of the International Conference on Performance Engineering (ICPE)*, 2019, pp. 1–11.
- [45] R. S. Kannan, L. Subramanian, A. Raju, J. Ahn, J. Mars, and L. Tang, "GrandSLam: Guaranteeing SLAs for Jobs in Microservices Execution Frameworks," in *Proc. of the 14th EuroSys Conference*, 2019, pp. 34:1–34:16.
- [46] N. C. Mendonça, D. Garlan, B. R. Schmerl, and J. Cámara, "Generality vs. reusability in architecture-based self-adaptation: The case for self-adaptive microservices," in *Proc. of the 12th European Conference on Software Architecture: Companion Proceedings (ECSA)*, 2018, pp. 18:1–18:6.
- [47] A. Bhattacharya and P. De, "A survey of adaptation techniques in computation offloading," *Journal of Network and Computer Applications*, vol. 78, pp. 97–115, 2017.
- [48] B. Zhang, X. Jin, S. Ratnasamy, J. Wawrzyniak, and E. A. Lee, "AW-Stream: adaptive wide-area streaming analytics," in *Proc. of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*, 2018, pp. 236–252.
- [49] S. Zhou, K. Lin, J. Na, C. Chuang, and C. Shih, "Supporting Service Adaptation in Fault Tolerant Internet of Things," in *Proc. of the 2015 IEEE 8th International Conference on Service-Oriented Computing and Applications (SOCA)*, 2015, pp. 65–72.
- [50] S. Mittal, "A Survey of Techniques for Approximate Computing," *ACM Computing Surveys*, vol. 48, no. 4, pp. 62:1–62:33, 2016.
- [51] O. Machidon, T. Fajfar, and V. Pejović, "Implementing Approximate Mobile Computing," in *Proc. of the 2020 Workshop on Approximate Computing Across the Stack (WAX)*, 2020, pp. 1–3.
- [52] T. Moreau, J. S. Miguel, M. Wyse, J. Bornholt, A. Alaghi, L. Ceze, N. D. E. Jerger, and A. Sampson, "A Taxonomy of General Purpose Approximate Computing Techniques," *Embedded Systems Letters*, vol. 10, no. 1, pp. 2–5, 2018.
- [53] V. Gupta, D. Mohapatra, S. P. Park, A. Raghunathan, and K. Roy, "IMPACT: imprecise adders for low-power approximate computing," in *Proc. of the 2011 International Symposium on Low Power Electronics and Design (ISLPED)*, 2011, pp. 409–414.
- [54] R. Ye, T. Wang, F. Yuan, R. Kumar, and Q. Xu, "On reconfiguration-oriented approximate adder design and its application," in *Proc. of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2013, pp. 48–54.
- [55] D. Mohapatra, V. K. Chippa, A. Raghunathan, and K. Roy, "Design of voltage-scalable meta-functions for approximate computing," in *Proc. of the 2011 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2011, pp. 950–955.
- [56] P. Guo, B. Hu, R. Li, and W. Hu, "FoggyCache: Cross-Device Approximate Computation Reuse," in *Proc. of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2018, pp. 19–34.
- [57] J. F. Pérez, R. Birke, and L. Y. Chen, "On the latency-accuracy tradeoff in approximate MapReduce jobs," in *Proc. of the 2017 IEEE Conference on Computer Communications (INFOCOM)*, 2017, pp. 1–9.
- [58] A. Agrawal, J. Choi, K. Gopalakrishnan, S. Gupta, R. Nair, J. Oh, D. A. Prener, S. Shukla, V. Srinivasan, and Z. Sura, "Approximate computing: Challenges and opportunities," in *Proc. of the IEEE International Conference on Rebooting Computing (ICRC)*, 2016, pp. 1–8.
- [59] Q. Zhang, F. Yuan, R. Ye, and Q. Xu, "ApproxIt: An Approximate Computing Framework for Iterative Methods," in *Proc. of the 51st*

Annual Design Automation Conference 2014 (DAC), 2014, pp. 97:1–97:6.

- [60] H. A. F. Almurib, T. N. Kumar, and F. Lombardi, “Approximate DCT Image Compression Using Inexact Computing,” *IEEE Transactions on Computers*, vol. 67, no. 2, pp. 149–159, 2018.
- [61] Q. Zhang, T. Wang, Y. Tian, F. Yuan, and Q. Xu, “ApproxANN: an approximate computing framework for artificial neural network,” in *Proc. of the 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2015, pp. 701–706.
- [62] C. Chen, J. Choi, K. Gopalakrishnan, V. Srinivasan, and S. Venkataramani, “Exploiting approximate computing for deep learning acceleration,” in *Proc. of the 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2018, pp. 821–826.
- [63] A. R. Zamani, I. Petri, J. D. Montes, O. F. Rana, and M. Parashar, “Edge-Supported Approximate Analysis for Long Running Computations,” in *Proc. of the 5th IEEE International Conference on Future Internet of Things and Cloud (FiCloud)*, 2017, pp. 321–328.
- [64] Z. Wen, D. L. Quoc, P. Bhatotia, R. Chen, and M. Lee, “ApproxIoT: Approximate Analytics for Edge Computing,” in *Proc. of the 38th IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2018, pp. 411–421.
- [65] D. Schäfer, J. Edinger, J. M. Paluska, S. VanSyckel, and C. Becker, “Tasklets: “Better than Best-Effort” Computing,” in *Proc. of the 25th International Conference on Computer Communication and Networks (ICCCN)*, 2016, pp. 1–11.
- [66] J. Edinger, D. Schäfer, M. Breitbach, and C. Becker, “Developing distributed computing applications with Tasklets,” in *Proc. of the 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2017, pp. 94–96.
- [67] V. Pejovic, “Towards Approximate Mobile Computing,” *GetMobile: Mobile Computing and Communications*, vol. 22, no. 4, pp. 9–12, 2018.
- [68] S. Sidiropoulos-Douskos, S. Misailovic, H. Hoffmann, and M. C. Rinard, “Managing performance vs. accuracy trade-offs with loop perforation,” in *Proc. of the SIGSOFT/FSE’11 19th ACM SIGSOFT Symposium on the Foundations of Software Engineering (FSE-19) and ESEC’11: 13th European Software Engineering Conference (ESEC-13)*, 2011, pp. 124–134.
- [69] D. Benavides, S. Segura, and A. R. Cortés, “Automated analysis of feature models 20 years later: A literature review,” *Information Systems*, vol. 35, no. 6, pp. 615–636, 2010.
- [70] J. Gedeon, M. Wagner, J. Heuschkel, L. Wang, and M. Mühlhäuser, “A Microservice Store for Efficient Edge Offloading,” in *Proc. of the IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.
- [71] J. Gedeon, M. Stein, L. Wang, and M. Mühlhäuser, “On Scalable In-Network Operator Placement for Edge Computing,” in *Proc. of the 27th International Conference on Computer Communication and Networks (ICCCN)*, 2018, pp. 1–9.
- [72] G. Gkioxari, J. Johnson, and J. Malik, “Mesh R-CNN,” in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9784–9794.
- [73] K. Kadir, M. K. Kamaruddin, H. Nasir, S. I. Safie, and Z. A. K. Bakti, “A comparative study between LBP and Haar-like features for Face Detection using OpenCV,” in *Proc. of the 4th International Conference on Engineering Technology and Technopreneuship (ICE2T)*, 2014, pp. 335–339.
- [74] J. Gedeon, M. Stein, J. Krisztinkovics, P. Felka, K. Keller, C. Meurisch, L. Wang, and M. Mühlhäuser, “From Cell Towers to Smart Street Lamps: Placing Cloudlets on Existing Urban Infrastructures,” in *Proc. of the IEEE/ACM Symposium on Edge Computing (SEC)*, 2018, pp. 187–202.