Microservice transition and its granularity problem: A systematic mapping study

Sara Hassan1 | Rami Bahsoon1 | Rick Kazman2,3

1School of Computer Science, University of Birmingham, West Midlands, UK
2Software Engineering Institute (SEI)/CMU, Pittsburgh, Pennsylvania, USA
3University of Hawaii, Honolulu, Hawaii, USA

Correspondence
Sara Hassan, Jeddah 2142, KSA.
Email: ssh195@cs.bham.ac.uk

Summary
Microservices have gained wide recognition and acceptance in software industries as an emerging architectural style for autonomic, scalable, and more reliable computing. The transition to microservices has been highly motivated by the need for better alignment of technical design decisions with improving value potentials of architectures. Despite microservices’ popularity, research still lacks disciplined understanding of transition and consensus on the principles and activities underlying that transition. In this paper, we report on a systematic mapping study that consolidates various views, approaches and activities that commonly assist in the transition to microservices. The study aims to provide a better understanding of the transition; it also contributes a working definition of the transition and technical activities underlying it. We term the transition and technical activities leading to microservice architectures as microservitization. We then shed light on a fundamental problem of microservitization: microservice granularity and reasoning about its adaptation as first-class entities. This study reviews state-of-the-art and -practice related to reasoning about microservice granularity; it reviews modeling approaches, aspects considered, guidelines and processes used to reason about microservice granularity. This study identifies opportunities for future research and development related to reasoning about microservice granularity.

KEYWORDS
design decision support, granularity, microservices, software economics, systematic mapping study

1 INTRODUCTION

Several industries have migrated their applications (or actively considering migrating) to microservices.1-3 The transition to microservices has not been purely driven by technical objectives; the transition requires careful alignment of the technical design decisions with the business ones. The ultimate objective of this alignment is to enhance utilities of the application’s software architecture and to improve its value potentials. For example, among the technical design decisions is isolating business functionalities into microservices that interact through standardized interfaces. Isolation motivated only by technical objectives can lead to aggressive decomposition of functionalities favouring service autonomy without considering its impact on value potentials. However, isolation motivated by technical and business objectives can be more
informed. It can enhance utilities such as autonomy, replaceability, and decentralized governance (among other utilities) to improve the microservice architecture’s ability in coping with operation, maintenance, and evolution uncertainties. Ultimately, this can also relate to improved maintenance costs and cost-effective quality of service (QoS) provision to end users; these are examples of improved value potentials in the architecture.

Due to the recency of microservices, they have a multitude of definitions; each definition captures different properties of microservices. Definitions mostly agree that the fundamental properties of microservices include enabling facilitated improvement of component characteristics—autonomy, replaceability, independent deployability—and of architectural characteristics—improved reliability, scalability, resilience to failure, availability, and evolvability. In essence, these definitions capture some drivers of the transition to microservices aimed at enhancing utility in the application’s software architecture. The utility enhanced through the transition can render benefits that can cross-cut architectural design, testing, maintenance and service management. For example, microservice autonomy allows the architects to easily locate, implement, and test necessary service amendments. Microservice replaceability allows architects to confidently and independently add and/or manage new business functionalities over the system lifetime.

The transition to microservices can help the application in better meeting its QoS requirements; this may consequently result into improved compliance with service level agreements for QoS, potential economics gains, and better alignment with the business objectives of microservice adopters. Because of their “micro” character, microservices can be mobilized to the benefit of several service-oriented applications that can require “lighterweight” processing (eg, mobile services and Internet-of-Things [IoT]).

Despite the industrial push toward microservices, there is no disciplined understanding of their transition nor consensus on the principles and activities underlying the transition. A disciplined understanding of the transition is of paramount importance to inform and/or to justify its technical activities by aligning them with their added value and cost. Currently however, the state-of-the-practice in microservice adoption lacks appropriate methods and techniques that can justify value added of technical design decisions. For example, the software architect can be equipped with mechanisms and tools that can enhance replaceability by standardizing the communication paradigms across microservices. Reasoning about the added value and possible cost becomes essential to justify this technical design decision regarding communication paradigms.

This paper is an attempt for a better understanding of the transition to microservices. It conducts a systematic mapping study to consolidate various views about microservices; it then uses the study results to contribute to a well-rounded working definition describing the transition and technical activities of the transition to microservice architectures. We term the transition and technical activities leading to microservice architectures as microservitization.

This working definition has explicitly considered a fundamental problem of microservitization: reasoning about the granularity of a microservice (ie, whether a microservice should be decomposed/merged further or not). A granularity level determines “the service size and the scope of functionality a service exposes (p. 426 of Reference 11).” Granularity adaptation entails merging or decomposing microservices thereby moving to a finer or more coarse grained granularity level.

Determining the granularity level too early in the software architecture’s lifetime can lead to problems in reasoning about microservices. This problem is of significance both in brownfield and greenfield development. In both fields, a suitable granularity level is paramount to inform choosing concrete services from a plethora of commercial-off-the-shelf (COTS) microservices. Conducting a systematic mapping study to examine microservice literature is therefore essential to inform the paramount exercise of reasoning about granularity.

In this paper, we consider that “a systematic mapping study allows the evidence in a domain to be plotted at a high level of granularity[p. 5 of Reference 14].” “This allows for the identification of evidence clusters and evidence deserts to direct the focus of future systematic reviews and to identify areas for more primary studies to be conducted[p. 5 of Reference 14].” Directing the focus of future systematic reviews aligns with our aforementioned objectives. Ultimately, our attempt at defining the transition to microservices will pave the way for future development and research related to microservice transition. Furthermore, understanding the problem of reasoning about microservice granularity allows identifying areas for future primary studies. Since the examined literature regarding microservices spanned a broad variety of aspects, we found a systematic mapping study to be suitable given the amount of reviewed literature.

In Section 2 we describe the steps we followed in the systematic mapping study. In Section 3 we report and briefly analyze our mapping study results. In Section 4 we use this analysis to: (i) present our working definition for the
transition to microservices (Section 4.1) and (ii) identify gaps in the state-of-the-art and -practice related to reasoning about microservice granularity (Section 4.2). Overall, the identified gaps motivate the need for:

- Microservice-specific modeling support potentially using an architecture definition language (ADL) that “treats microservice boundaries as adaptable first-class entities (p. 2 of Reference 15)” thereby facilitating runtime analysis of microservice granularity in a systematic architecture-oriented manner.\(^{15}\)
- A dynamic architectural evaluation approach that captures two dimensions under uncertainty at runtime: added value to be introduced and cost to be incurred if the granularity of microservice architectures is adapted.
- Effective decision support that should systematically guide the software architects toward suitable granularity adaptation strategies at runtime or suggest revisiting their expectations of the architecture’s runtime environment.

In Section 6 we compare and contrast related literature reviews, studies and surveys in the field of microservices against our systematic mapping study. In Section 5 we reflect on threats to the validity of our study. In Section 7 we summarize contributions that are not directly linked to microservices but can be relevant to their emergence and development. In Section 8 we conclude by summarizing the results of our systematic mapping study.

## 2 | SYSTEMATIC MAPPING STUDY PROCESS

The process we follow in our mapping study is inspired by guidelines from Reference 14. In the subsections below, we describe our application of each stage of this process.

### 2.1 | Research questions

Overall, this paper conducts a systematic mapping study to address the following objectives:

- **Objective 1**: providing a better understanding of the transition to microservices—we consolidate various views (industrial, research/academic) of the principles, methods, and techniques that are commonly adopted to assist the transition to microservices. This consolidation allows us to reach a working definition for the transition to microservices; we term this transition *microservitization*.

- **Objective 2**: understanding a fundamental problem of the transition to microservices related to reasoning about their granularity—we review state-of-the-art and -practice related to reasoning about microservice granularity. This review allows us to understand the state-of-the-art and -practice in the modelling approaches, aspects considered, guidelines and processes used to reason about microservice granularity.

Given these objectives, we reify them into the following research questions. Along with each question we outline the rationale behind it.

**Objective 1: Providing a better understanding of the transition to microservices**

- What are the activities undertaken to adopt microservices? This question helps understand the principles, methods and techniques of the transition to microservices by digesting experiences of microservice adopters in industry and in academic research.

**Objective 2: Understanding a fundamental problem of the transition to microservices related to reasoning about their granularity**

- What are the modeling approaches used to define the granularity of a microservice? The aim of this question is to identify the support provided by microservice-specific models for reasoning about granularity and to investigate how systematic (ie, standardized and methodological) these models are.
- Which quality attributes are considered when reasoning about microservice granularity and how are they captured? The aim of this question is to elicit the possible trade-offs which software architects need to balance when reasoning about microservice granularity and/or how these trade-offs can be captured objectively.
• How is reasoning about microservice granularity described? The aim of this question is to explore the state-of-the-art regarding triggers and steps of microservice granularity adaptation and their suitability to the dynamic microservice environment.

2.2 Search strategy

The terms used when searching for English publications were “microservice” and “micro-service”; Google Scholar, Association for Computing Machinery Digital Library, Institute of Electrical and Electronics Engineers (IEEE) Xplore, ScienceDirect, SpringerLink, and Wiley InterScience were used during this search. Our scope of publications includes journals, theses, books, conferences, workshops, blog articles, presentations, and videos. For academic publications, snowballing was applied to further extract relevant literature. Nonacademic publications were included since most industrial experiences regarding microservices were published in these forms. We made our best effort, however, to only include articles that either transcribe the view of adopters or ones where the author is the opinion holder. We believe answering the research questions above comprehensively requires examining both academic and nonacademic experiences with microservices. This broad scope also aligns with a property of systematic mapping studies—aiming for broad coverage rather than narrow focus. Our search was restricted to publications between January 2013 and April 2020, since the microservice trend had not emerged prior to that; nonacademic literature started to appear in 2013 while peer-reviewed publications started to appear in 2014. Meta-data of the search results was maintained using a tool called “Publish or Perish.”

2.3 Selection of primary studies

Initially, the search results were examined for relevance according to inclusion and exclusion criteria below. Each research question elicited in Section 2.1 has corresponding inclusion/exclusion criteria (their structure is inspired by Reference 20). Along with each criterion is the rationale behind it. The rationale are presented in italics.

What activities are undertaken to adopt microservices?

**Inclusion Criteria:**

- Publications generically presenting the challenges of adopting microservices *since they can be used to infer the activities comprising the transition to microservices.*

- Case studies of adopting microservices *are used to complement and verify the activities in generic publications.*

- Publications comparing specific development tools in the microservice industry *because they can be used to infer activities comprising the transition.*

**Exclusion Criteria:**

- Publications without any reference to microservices. *Including such publications would confuse rather than clarify understanding the transition to microservices. This is against Objective 1 of our study.*

- Publications that refer to servitization in the business not the software context *since we are only concerned about the activities of shifting a software system from another architectural style to microservices.*

What modeling approaches are used to define the granularity of a microservice?

**Inclusion Criteria:**

- Publications defining formal notations/diagrams for modeling microservices. *Such publications can be used to assess how systematic the state-of-the-art is for modeling microservice granularity.*

- Proposals of modeling concepts for microservices. *Even when unverified, a proposed modeling concept can provide an insight for the building units of reasoning about microservice granularity.*

- Publications presenting industrial case studies for modeling microservices. *Such publications would verify and illustrate the expressiveness of proposed modeling concepts to microservice granularity.*
Exclusion Criteria:

- Papers which provide binding and reconfiguration solutions for service-oriented architectures (SOAs)/web-services/mobile services only are excluded since they do not capture the properties specific to microservices, hence they are not suited for reasoning about microservice-specific decision problems (in this case microservice granularity).
- Papers which provide modeling approaches for SOAs/web-services/mobile services are excluded since they do not capture properties specific to microservices, hence they are not suited for reasoning about microservice-specific decision problems (in this case microservice granularity).

Which quality attributes are considered when reasoning about microservice granularity and how are they captured?

Inclusion Criteria:

- Publications presenting metrics used when reasoning about microservice granularity. Such publications would help assess how objectively the trade-offs affecting microservice granularity are captured in academia and/or industry.
- Case studies involving the quality drivers considered when reasoning about granularity. Such publications would realistically capture the significance of specific trade-offs when reasoning about granularity adaptation.
- Publications focused on vendor-specific comparisons between platforms supporting reasoning about microservice granularity. This can help derive the quality attributes and metrics considered when reasoning about granularity.

Exclusion Criteria:

- Case studies that do not explicitly relate a challenge to its impact on microservice granularity. Since case studies report concrete challenges and trade-offs impacting them, it is unreasonable to claim an impact of a reported trade-off on granularity if that is not reported explicitly in a case study.

How is reasoning about microservice granularity described?

Inclusion Criteria:

- Publications including guidelines for reasoning about microservice granularity. This helps to identify the state-of-the-art regarding triggers and/or steps for granularity adaptation.
- Publications showing a sequence of when and how granularity is reasoned about in the application’s lifecycle. These publications help assess how much the state-of-the-practice considers dynamicity in microservice environments when reasoning about granularity adaptation.

Exclusion Criteria:

- Publications that provide generic best practices for the granularity of applications with no reference to microservices (eg, related to web services, SOA, mobile services). Such best practices are not targeted specifically at microservices, so it would not be reasonable to use them in the context of microservice granularity.

2.4 Keywords and classification

“The purpose of this stage is to classify papers with sufficient detail to answer the broad research questions and identify papers for later reviews without being a time-consuming task [p44 of Reference 14].” Here we classify the included publications according to two classification frameworks. Initially, we classify them according to their research approaches. We elicit these approaches from the seminal guidelines in Reference 21.

The second classification framework entails categorizing the publications under categories derived from our research questions of concern. A publication belongs in a category if it contains any of the corresponding keywords identified in Table 1. The identification of the keywords is inspired by a microservice-specific systematic mapping study and refined iteratively as more publications were examined. It is worth noting that for all the included publications, we manually
<table>
<thead>
<tr>
<th>Research question</th>
<th>Category</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>What activities are undertaken to adopt microservices?</td>
<td>Architectural design</td>
<td>Architectural style, communication mechanism, boundaries, orchestration, service choreography, service registration, service discovery, design patterns, proxy, bulkhead, circuit breaker, router, routing</td>
</tr>
<tr>
<td></td>
<td>Organization</td>
<td>Conway's law, decentralized governance, cross-functional teams, hierarchical teams</td>
</tr>
<tr>
<td></td>
<td>Operation</td>
<td>Devops, NoOps, configuration settings, operation</td>
</tr>
<tr>
<td></td>
<td>Deployment</td>
<td>Continuous integration, CICD, continuous deployment, deployment pipeline, automated deployment, virtualisation, hypervisors, containerisation, configuration provider</td>
</tr>
<tr>
<td></td>
<td>Development</td>
<td>Heterogeneous tools, agile development, extreme programming</td>
</tr>
<tr>
<td></td>
<td>Monitoring</td>
<td>Regression unit testing, health monitoring, cluster monitoring, troubleshooting, debugging, failure</td>
</tr>
<tr>
<td></td>
<td>Logging</td>
<td>Central logging, decentralized logging, profiling, tracing</td>
</tr>
<tr>
<td>What modeling approaches are used to define the granularity of a microservice?</td>
<td>Structural</td>
<td>Domain-driven, classes, instances, resources, components, message format, data item, topology, service dependency, nodes, type definition</td>
</tr>
<tr>
<td></td>
<td>Behavioral</td>
<td>Event flow, message stream, activity flow, communication flow, event triggers, execution timeline, use cases</td>
</tr>
<tr>
<td>Which quality attributes are considered when reasoning about microservice granularity and how are they captured?</td>
<td>Performance</td>
<td>QoS, efficiency, service contracts, SLA, performance, response time, throughput, performance bottleneck, invocation duration, transaction duration, customer value</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
<td>Fault tolerance, disaster recovery, single point of failure, resilience, robustness, failure rate, error rate</td>
</tr>
<tr>
<td></td>
<td>Scalability</td>
<td>Auto-scale, scalable, scaling, load balancing, load distribution, completed transactions per second</td>
</tr>
<tr>
<td></td>
<td>Maintainability</td>
<td>Maintainable, changeable, maintenance cost, maintenance overhead, adaptability, changeability, effort cost, expandability, dynamicity</td>
</tr>
<tr>
<td></td>
<td>Complexity</td>
<td>Communication overhead, complexity cost, development cost, tight coupling, low cohesion</td>
</tr>
<tr>
<td>How is reasoning about microservice granularity described?</td>
<td>Guidelines</td>
<td>Two-pizza team, lines of code, half-life, agile manifesto, one task, single responsibility, fine-grained functionality, separation of business concerns, high cohesion, loose coupling</td>
</tr>
<tr>
<td></td>
<td>Processes</td>
<td>Iterative, strangler pattern</td>
</tr>
</tbody>
</table>

Abbreviations: CICD, continuous integration and continuous delivery; QoS, quality of service; SLA, service level agreement.

categorized them to ensure that synonyms or partial matches of these keywords are accurately handled. We justify the keywords (italicized below) under each category as follows:

*What activities were undertaken to adopt microservices?*

- **Architectural design:** publications in this category are concerned with all the technical activities which comprise adopting microservices. We use this category to verify whether or not microservice adopters call microservices an architectural style and/or design pattern. Choosing the generic communication paradigm of the architecture (eg, orchestration, choreography) and the more concrete message exchange pattern (eg, using a router/proxy) are also among the technical activities of microservice adoption. In addition, the boundaries of the system and individual components of the system is a technical design decision when moving to microservices. The fault tolerance mechanisms of the system also need to be identified (eg, bulkhead and circuit breaker patterns). Because microservice applications are highly distributed and scalable, two additional technical decisions are critical in microservitization: service registration and service discovery.
Organization: in this category we are concerned with the organisational impact of adopting microservices. The state-of-the-practice in the microservice industry is to motivate decentralised governance (ie, holding the responsibility of building and running) through microservitization. The three most common means of achieving that is through following Conway’s law, cross-functional teams, or hierarchical teams. Conway’s law states "organisations which design systems … are constrained to produce designs which are copies of the communication structures of these organisations."

Operation: publications in this category are concerned with identifying how the system will be governed post-deployment. This includes defining who is responsible for this governance to begin with. The state-of-the-practice in the microservice industry is two alternative approaches: DevOps where the governance is shared across a development team and an operations team; and NoOps, where the governance is fully the responsibility of the development team. In addition, operational activities include defining the configuration settings and configuration provider which the governor of the microservice application need to follow or in different runtime situations.

Deployment: this category includes publications that define approaches for managing the deployment pipeline of a microservice application. The state-of-the-practice in the microservice industry is two alternative approaches: continuous integration and continuous delivery (abbreviated as CICD), and automated deployment. “Continuous integration is a coding philosophy and set of practices that drive development teams to implement small changes and check in code to version control repositories frequently.” “Continuous delivery picks up where continuous integration ends. Continuous delivery (CD) automates the delivery of applications to selected infrastructure environments.” “Deployment automation allows applications to be deployed across the various environments used in the deployment process, as well as the final production environments.” Common deployment automation tools in the microservice field include Kubernetes and Istio. Regardless of the deployment pipeline, the host on which this pipeline is enforced is another critical decision in this category. The three most common approaches for deploying microservices are virtualization, using hypervisors, and/or containerization. Cross-cutting these approaches are deployment activities that can be motivated by different quality attributes (eg, load-balancing and auto-scaling motivated by scalability). To answer the question of identifying which quality attributes are considered when reasoning about microservice granularity we find it more be-fitting to categorize these activities as keywords under their related quality attributes.

Development: publications in this category describe how the transition to microservices impacts software development practices. The state-of-the-practice in microservice development is adopting extreme programming and agile practices using heterogenous tools (eg, Springboot, Fabric8).

Monitoring: publications in this category are concerned with identifying the alternative rationales of runtime monitoring (eg, health, cluster) of microservice application and the alternative approaches which support monitoring (eg, regression unit testing, troubleshooting, debugging, and failure identification).

Logging: in this category we are concerned with where the monitoring results are to be stored. Logging is alternatively called profiling or tracing. The two alternative approaches of logging in the microservice industry as we have examined are central or decentralized logging.

What modeling approaches are used to define the granularity of a microservice?

Structural: publications in this category are concerned about contributions that capture the structure (alternatively called topology) of the microservice architecture. Depending on the nature of the contribution (eg, domain-driven), the units of this structure differ (eg, classes, instances, resources, components, data items, messages, and/or nodes). Modeling these units includes capturing the dependencies across those units and their types.

Behavioral: alternative to the approach above, publications in this category capture the sequence of actions (ie, events, messages, activities, communication, execution steps, and/or use cases) of a microservice application. The granularity of a microservice in this case would be defined in terms of which actions a single microservice is responsible for.

Which quality attributes are considered when reasoning about microservice granularity and how are they captured?

Performance: wherever the main driving force of reasoning about granularity is performance (alternatively called QoS, long-term value, efficiency or customer value), the publication is put under this category. The metrics for capturing performance objectively include response time, throughput, invocation duration, identifying the performance
bottlenecks, and/or transaction duration. Thresholds on these metrics are captured in service contracts, service level agreements (abbreviated as SLAs). We subsume service contracts as “an agreement between the a consumer and provider service about the format of data that they transfer between each other. Normally, the format of the contract is defined by the consumer and shared with the corresponding provider. Afterwards, tests are being implemented in order to verify that the contract is being kept.30"

- Reliability: publications that imply or explicitly focus on enhancing reliability as the primary driving force when reasoning about granularity are included in this category. Reliability entails exhibiting fault tolerance, disaster recovery, resilience, eliminating single points of failure, and/or robustness. Reliability is captured objectively in terms of failure/error rate.

- Scalability: publications that reason about granularity in terms of how it enhances the scalability of the architecture are included in this category. Exhibiting scalability entails employing strategies such as auto-scaling, load balancing, and/or load distribution. Measuring scalability objectively can be done by looking at the number of completed transactions per second (or per unit of time more generally).

- Maintainability: exhibiting maintainability entails exhibiting adaptability, changeability, expandability, and/or dynam- icity. These properties are objectively captured in terms of maintenance cost, maintenance overhead, and/or effort cost. Publications concerned with reasoning about granularity in terms of enhancing maintainability are included in this category.

- Complexity: minimizing complexity entails following two crucial design principles: tight coupling and/or low cohesion. This is measured in terms of communication overhead, complexity cost, and/or development cost. Publications where reasoning about granularity considers and/or measures its complexity are included in this category.

How is reasoning about microservice granularity described?

- Guidelines: publications under this category provide decision-making strategies for reasoning about granularity regardless of the steps of applying these strategies. The keywords under this category capture the state-of-the-practice guidelines.

- Processes: publications under this category are more elaborate in the sense that they enrich the granularity adaptation strategies with a sequence for applying them. The keywords under this category capture the alternative state-of-the-practice processes encountered for reasoning about microservice granularity.

2.5 | Data synthesis

Really Simple Syndication (RSS) feeds and manual search were used to obtain publications complying with the strategy defined in Section 2.2. The results were then manually examined for inclusion and categorised according to the criteria and frameworks described above (Sections 2.3 and 2.4, respectively).

3 | RESULT REPORTING AND ANALYSIS

In this section we present graphs summarizing distributions of the included publications along the categories described in Table 1. For each graph, we discuss how it helps serve the objectives outlined in Section 1.

3.1 | Publication distribution overview

A total of 877 publications met the inclusion criteria in Section 2.3. Table 2 lists representative examples of the included publications categorized according to Table 1. Figure 1 shows the overall distribution of the publications according to the publication type.

As justified in Section 2, we widened the scope of our study to include both academic and industrial publications. Broadly, we consider a publication type to be academic if it has gone through editing or peer-revision (the red bars in Figure 1); about 63% of the included publications. On the other hand, nonacademic publications account for about 36%
TABLE 2  Representative examples of publications included in the systematic mapping study

<table>
<thead>
<tr>
<th>Research question</th>
<th>Category</th>
<th>Representative examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>What activities are undertaken to adopt microservices?</td>
<td>Architectural design</td>
<td>31-44</td>
</tr>
<tr>
<td></td>
<td>Organization</td>
<td>45-54</td>
</tr>
<tr>
<td></td>
<td>Operation</td>
<td>48,55-67</td>
</tr>
<tr>
<td></td>
<td>Deployment</td>
<td>68-79</td>
</tr>
<tr>
<td></td>
<td>Development</td>
<td>80-92</td>
</tr>
<tr>
<td></td>
<td>Monitoring</td>
<td>93-99</td>
</tr>
<tr>
<td></td>
<td>Logging</td>
<td>100-102</td>
</tr>
<tr>
<td>What modeling approaches are used to define the granularity of a microservice?</td>
<td>Structural</td>
<td>29,103-112</td>
</tr>
<tr>
<td></td>
<td>Behavioral</td>
<td>8,113-117</td>
</tr>
<tr>
<td>Which quality attributes are considered when reasoning about microservice</td>
<td>Performance</td>
<td>118-124</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
<td>125-130</td>
</tr>
<tr>
<td></td>
<td>Scalability</td>
<td>131-135</td>
</tr>
<tr>
<td></td>
<td>Maintainability</td>
<td>101,136-138</td>
</tr>
<tr>
<td></td>
<td>Complexity</td>
<td>139-144</td>
</tr>
<tr>
<td>How is reasoning about microservice granularity described?</td>
<td>Guidelines</td>
<td>145-155</td>
</tr>
<tr>
<td></td>
<td>Processes</td>
<td>110,156-164</td>
</tr>
</tbody>
</table>

FIGURE 1  Publications between January 2013 and April 2020 as per the search strategy defined in Section 2.2 included as per criteria from Section 2.3 classified according to their publication type; the red bars are the publications we consider as academic and the blue bars are those we consider nonacademic. [Colour figure can be viewed at wileyonlinelibrary.com]

of the total. Although the majority of publications are academic, nonacademic literature still comprises a significant percentage. Had we excluded these publications, attempting a definition for microservitization (Objective 1 in Section 1) would have been biased. The exclusion of nonacademic sources would lead to missing relevant keywords related to each category and the coverage of industrial opinions and experiences related to microservice adoption would be narrower.

We further classify peer-reviewed publications according to a highly cited paper classification framework which targets IEEE papers. This framework classifies papers according to their research approach; a brief description of each approach is presented in Section 2. This framework has been applied before in the context of microservices, so we consider it a neat fit for our study. To match the target context of the framework, we only apply it to peer-reviewed publications (Figure 2).

Solution proposals by far comprise the largest number of peer-reviewed publications. Solution proposals present novel, significant techniques without a full-blown validation. A proof-of-concept may be offered in solution proposals by
means of a small example, a sound argument, or by some other means. \textsuperscript{21} Therefore, the microservices trend is a thriving field for novelty but it is still lacking maturity. Validation research publications — which thoroughly investigate solution proposals \textsuperscript{21} — only amounted to 73 publications, which further proves the lack of maturity in the field. The large difference between the solutions proposals and validation research publications proves the need for disciplining the transition to microservices. The following subsections further discuss this need then focus on one of the fundamental problems of the transition — reasoning about microservice granularity.

3.2 \hspace{1em} \textbf{Objective 1: Providing a better understanding of the transition to microservices}

Figure 3 classifies the publications which include keywords related to: \textit{what are the activities undertaken to adopt microservices?} Architectural design and managing deployment are the most popular activities undertaken when adopting microservices. Therefore, we infer they are crucial activities in the transition to microservices. Nevertheless, there is a significant number of publications in the other categories of Figure 3. The variation in number of publications across categories of Figure 3 implies there is no consensus in describing the transition to microservices. This leaves room for us to contribute the microservitization term which attempts to provide a better understanding of the transition to microservices.

3.3 \hspace{1em} \textbf{Objective 2: Understanding the microservice granularity problem}

One of the fundamental problems of the transition to microservices is finalising their level of granularity.\textsuperscript{165} ADL classification frameworks\textsuperscript{166} indicate that structural (or “topological[p26 of Reference 166]”) as well behavioral aspects of an
architecture need to be modeled. Figure 4 helps to clearly identify the state-of-the-practice in modeling microservices; to answer what are the modeling approaches used to define the granularity of a microservice? The structural modeling approaches proposed for microservices are almost double the behavioral approaches. Structural approaches capture the topology and/or dependencies across building units of the microservice architecture. On the other hand, behavioral approaches capture the actions of these units in several runtime scenarios in which the modelled microservice operates. Therefore, a systematic, architecture-oriented modeling approach for microservice architectures which facilitates reasoning about granularity needs to capture the architecture’s structural and behavioral aspects. The difference in numbers between structural- and behavioral-oriented publications in Figure 4 indicates that there is a lack in modeling approaches that capture both aspects of a microservice architecture.

Figure 5 classifies publications according to the quality attributes they aim to optimise when reasoning about microservice granularity; to answer which quality attributes are considered when reasoning about microservice granularity and how are they captured? In other words, they can be the most common means to introduce utility through cost-effective microservitization. Scalability is the most common quality considered in the examined literature. This is reasonable given the dynamic, large-scale environment in which microservices operate. Therefore, we infer that scalability can introduce added value to most microservice architectures. Relatively few publications have considered complexity/cost when reasoning about microservice granularity. Therefore, there is room for contributing to dynamic decision support which objectively considers both the potential value and cost of decisions related to microservice granularity (ie, adapting granularity by decomposing or merging microservices).

Figure 6 classifies proposed approaches for reasoning about microservice granularity according to how they are described; this answers how is reasoning about microservice granularity described? Most of the proposed approaches are ad-hoc guidelines that could be applied differently at various points of the microservice application’s lifecycle. However, effective decision support for microservice granularity should comprise a clear sequence of distinct steps to be triggered under clear conditions. We have not found such effective support even in the 33 publications proposing a process to reason about microservice granularity. Therefore, we infer that there is a lack in effective decision support for reasoning about microservice granularity in terms of clear adaptation steps and triggers.
### 4 | ADDRESSING SYSTEMATIC MAPPING STUDY OBJECTIVES

In Section 4.1 we contribute to an empirically grounded working definition for the transition to microservices; we call this transition microservitization. In Section 4.2, we identify the gaps in state-of-the-art and -practice related to reasoning about microservice granularity (inferred from Section 3) and discuss how they can be addressed.

#### 4.1 | Objective 1: Providing a Better Understanding of the Transition to Microservices

We have not found in the surveyed publications an empirically grounded definition which characterises the transition to microservices, but rather several conflicting, informal attempts coming from industry and academia. Table 3 analyzes these attempts in terms of whether or not they explicitly include the activities derived from the microservice-specific systematic mapping study\(^{13}\) and described in Section 2.4. It is worth noting that the attempts analyzed in this table are just a fraction of the publications summarized in Figure 3. In Table 3 we focus on the explicit attempts to define the transition.

| TABLE 3 | Analyzing publications that attempt to define the transition to microservices included in the systematic mapping study; a T means the publication includes the activity in the corresponding column |
|-----------------------------------------------|
| **Publication** | **Architectural design** | **Organization** | **Operation** | **Deployment** | **Development** | **Monitoring** | **Logging** |
| 165 | T | T | T |  |
| 168, 168 | T |  |
| 169, 169 | T | T | T |  |
| 170, 170 | T | T | T | T |  |
| 171, 171 | T | T | T | T |  |
| 172, 172 | T |  |
| 173 | T |  |
| 174 | T | T |  |
| 175 | T | T | T |  |
| 176 | T | T | T | T | T |  |
| 177 | T |  |
| 18 | T | T | T |  |
| 12 | T | T |  |
| 178 | T | T | T |  |
| 179 | T | T |  |
| 153 | T |  |
to microservices, while Figure 3 includes both explicit attempts to define the transition and case studies of microservice adoption which do not explicitly attempt to define the transition.

Observing Table 3, we introduce a definition for the transition to microservices (adapted from the activities included in Reference 13) which we call microservitization to cover all the relevant activities of the transition to microservices. This definition is adapted from activities included in Reference 13 and inspired by a trending concept in business and manufacturing domains — servitization.

In the manufacturing and business domains, servitization is seen as a paradigm shift entailing “manufacturers growing their revenues and profits through services” rather than tangible functional products. A service in this context is any feature that helps the business to (i) “really make money” and (ii) deliver new outcomes to customers. Examples of services in servitization include software applications, customer support, and self-service capabilities.

Servitization “embraces business model innovation, organisational change, and new technology adoption. Services exist in various forms, and represent differing values to both the customer and provider.” To benefit from these values, servitization involves developing new relationships with customers, innovating customer value propositions, forming new value chain relationships, and adapting business models. The key to successful servitization is “choosing the right technology, picking the appropriate moment to invest, and ensuring successful implementation” which aligns with the business objectives.

We liken the transition to microservices to servitization because of the following resemblances:

• Servitization is driven by delivering new outcomes to customers which can translate into revenues and profits. Similarly, the transition to microservices is driven by improving QoS provision to end users and translating that into an economic gain.

• Servitization entails embracing innovation in the manufacturing activities (eg, building business models and defining customer value propositions) are carried out. Similarly, the transition to microservices entails a dramatic change (motivated by value creation) to the way technical activities manifested in the software architecture are carried out.

Therefore, we define microservitization as a form of servitization where “services/components are transformed into microservices — a more fine-grained and autonomic form of services” — to introduce added value to the architecture. “Microservitization is also an example of a paradigm shift since it involves dramatically changing how the following technical activities are carried out to align them with a microservice adopter’s business objectives:

• Architectural design: microservitization introduces the following critical architectural design activities:
  - Choosing light-weight communication mechanisms: microservitization can increase the distribution of functionalities across the architecture thereby introducing extra communication calls between microservices. Therefore, rather than relying on enterprise service buses (which are the state-of-the-practice in SOA architectures), more light-weight mechanisms such as pipes and filters, event-based queues, and correlation identifiers are critical to avoid very high communication costs in microservice architectures. The large number of microservices can lead to a large volume of message exchange and hence high communication costs.
  - Reasoning about microservice granularity levels: a suitable granularity level is paramount to inform buying COTS concrete services or developing them in-house. Choosing these services correctly is critical to introducing added value to the microservice architecture.
  - Adopting fault tolerance design patterns: although striving for fault tolerance is a best practice in any architecture, investing in fault tolerance design patterns is all the more critical to microservitization. The criticality is due to the scale of industries adopting microservices (eg, retail, entertainment) where microservices span different continents with a wide variety of end users. Microservice-specific fault tolerance design patterns include circuit breakers and bulkheads.
  - Incorporating microservice registration and discovery mechanisms: microservices are typically developed, deployed and replaced at a very quick rate. Therefore, it is critical to incorporate robust registry and discovery mechanisms in microservice architectures to ensure an up to date record of the currently “alive” microservices.

• Managing the organizational hierarchy: microservitization has a direct impact on the organizational hierarchy. In particular, the autonomy and independent deployability enhanced in the architecture through microservitization facilitate decentralised governance by breaking “silos” (based around strict separation of job roles) in the organization.

• Operation: microservitization aligns operation management with breaking organizational “silos.” DevOps and NoOps are among the state-of-the-practice operation management approaches in microservitization; they involve “a set of
practices intended to reduce the time between committing a change to a system and the change being placed into normal production, while ensuring high quality. Decentralised operation management can reduce the risk of bottlenecks that can materialise into economic losses to the microservice adopter. Reducing this risk is conditional upon development operation teams adhering to SLAs between them.

- **Deployment**: microservitization introduces a critical challenge of determining the hosts on which a deployment pipeline is implemented. This is critical to balancing between the added value of microservitization and cost that can be introduced by a deployment pipeline implementation choice (eg, physical link installation, server rental and maintenance costs). Virtualization and containerization are among the common deployment pipeline implementation choices. They can enable swift auto-scaling the microservice architecture in response to changes in its runtime environment; this can materialize into economic gains for competitive microservice adopters.

- **Development**: unlike other software development paradigms, microservitization enables freedom in choosing development tools which can in turn introduce more added value to the architecture. This freedom is a bi-product of the decentralised governance enabled by microservitization. It is worth noting that communication and knowledge sharing across teams using different development tools needs to be carefully managed to ensure this freedom actually introduces added value.

- **Monitoring**: microservitization requires much more robust, decentralised and customisable monitoring than that of classical SOAs due to the heterogeneity of tools used to develop microservices and scale at which they typically operate. These requirements are critical to cater for the heterogeneity, scale, and dynamism of microservice architectures.

- **Logging**: maintaining logs of the monitoring data needs to be more customisable and distributable than logging classical SOAs due to the heterogeneity of tools used to develop microservices and scale at which they typically operate. This requirements are aligned with the aforementioned monitoring challenges introduced by microservitization.

4.2 | **Objective 2: Understanding the microservice granularity problem**

Based on our definition above, microservitization introduces the challenge of reasoning about the suitable granularity level of a microservice.

To formalize the microservice granularity problem, the gap we identified is the lack of an architecture-oriented modeling approach that captures a microservice’s granularity behaviour, thereby supporting runtime analysis of this behavior. The approach should treat microservice boundaries as the primitives for formulating the microservice granularity decision problem and actuators of microservice granularity adaptation decisions. These decisions include merging multiple microservices into a single boundary and decomposing a microservice into multiple ones encapsulated by multiple boundaries. In other words, this approach should treat microservice boundaries as adaptable first-class entities to ensure that both the structural and behavioral aspects of the microservice architecture are captured; we contribute to this approach in Reference 15. While conducting our study, we have seen architecture-oriented modeling approaches that treat the notion of boundaries statically (eg, Reference 189), or provide support for adaptability but without explicitly capturing the notion of adaptable boundaries (eg, Reference 190). Because the role of microservices is to encapsulate functionality, it is intuitive to use boundaries as adaptable first-class entities in the modelling approach. By contrast, if the decision problem was to determine the optimal physical infrastructure to host the microservice, for example, the adaptable first-class entity would be different (eg, microservice configuration variables).

To reason about microservice granularity objectively and dynamically, the gap we identified is the lack for an architectural evaluation approach that captures two aspects explicitly: added value to be introduced and cost to be incurred by pursuing granularity adaptation.

To effectively support reasoning about granularity adaptation, the gap we identified is the need for a decision support tool for reasoning about this problem at runtime. It is seen as a runtime problem since the suitable granularity level highly depends on the current scenario in which the microservice architecture is operating. For example, if a certain functionality in a microservice-based application is continuously receiving a large volume of requests at runtime, it makes sense to decompose this functionality in a separate microservice to manage its load separately. On the other hand, if two microservices are continuously communicating across a network at runtime causing latency, then merging these microservices is sensible to help reduce such latency. Overall, uncertainties related to the expected environment and behavior of the microservice architecture can not be fully captured at design time. Therefore, a runtime decision support tool is necessary to track and analyze this uncertainty. The tool should systematically guide the software architects
toward suitable granularity adaptation strategies at runtime or suggest revisiting their expectations of the microservice runtime environment. Each candidate granularity adaptation strategy must be systematically described as a sequence of merging/decomposition steps accompanied by triggers on them. Moreover, the tool’s suggestions need to be justified objectively while leaving the final decision to the architects for adopting the suggested strategies; approaches such as Reference 192 can inspire the design of this tool.

5 | THREATS TO VALIDITY

In this subsection we acknowledge the threats to validity in the process we use in our systematic mapping study as well as the application of each stage. Threats to validity are “influences that may limit our ability to interpret or draw conclusions from the study’s data (p351 of Reference 193)”.

When defining our search strategy in Section 2.2, we considered blog articles, presentations, and videos as the means of reporting first-hand industrial experiences with microservice adoption. We acknowledge that an alternative means could be interviews with microservice adopters in the industry. However, published blog articles, presentations and videos are arguably more trusted since they present a more responsible and objective view than interviews. Though they can enrich the study with diverse opinions, interviews tend to suffer from bias, subjectivity, and irresponsible answers. Moreover, our search strategy yields publications that explicitly mention microservices. Nevertheless, we acknowledge that there are publications prior without direct mention of microservices that could be relevant to their emergence and thereby to our research (eg, web service composition and agile development). We attempt to address this threat briefly in Section 7.

We acknowledge that the inclusion and exclusion criteria might have led to missing contributions that can inspire the microservices field. However, since our study was motivated by studying the state-of-the-art and -practice in the microservices trend we based the criteria on publications which already have a link to microservices. Nevertheless, we briefly outline the research areas that can inspire the microservice trend in Section 7. On a more technical level, we excluded publications that could not be translated to English which might have affected the study results.

We iteratively built Table 1 to include all the relevant keywords and made our best effort to justify them (Section 2). However, we acknowledge that some keywords might have been missed related to each research question.

When extracting videos and presentation for inclusion in the study results, we made our best effort to include videos whose content contained keywords from each category. The keywords are either mentioned by the speaker in the video or in the slides presented in the presentation. We acknowledge, however, that this might have biased the study results and that in the future transcription of the videos/presentations would be a more accurate means of determining their relevance.

When categorizing publications according to Table 1, we made our best effort to considers synonyms of the keywords in each category. However, since we categorized the publications by manual examination, we acknowledge that their distributions might have been skewed. In particular, we acknowledge that subjective interpretation of keywords might need to be complemented with more systematic approaches of categorizing the yielded publications. For example, the context in which a keyword or a fragment is mentioned needs to be considered before putting each publication under a certain category. A more systematic categorisation of the publications can ensure the reproducibility of our results.

We acknowledge that skewed distributions might lead to biassed inferences regarding gaps in the literature related to each objective of our study. For example, the variation in numbers of publications across categories in Figure 3 might be due to the inadequacies in keywords under each category. It might also indicate interest in architectural design across practitioners (ie, in nonacademic publications); this might not be as accurate for academic publications. Nevertheless, we argue that our mapping study results give a strong insight into the microservice trend and opens directions for more detailed research down each direction. For example, a systematic mapping study can be conducted which focusses specifically on microservice architectural design and/or development—the most dense categories in Figure 3.

6 | RELATED STUDIES

Motivated by disciplining the understanding of microservices, several studies have been conducted to examine the existing literature in this young yet trending field. They analyze existing literature with a variety of focuses. In this section, we compare and contrast the examined studies against the systematic study we conducted.
The closest to our study are References 13, 37, 195, 196 since they both adopt a systematic mapping study process when examining the literature. However, the motivations in these studies are different from ours. In Reference 13, for example, the research questions motivating the study are related to challenges, modeling approaches and quality attributes considered when adopting microservices. These questions do overlap partially with both objectives of our study but they do not focus on reasoning about microservice granularity as we do in our study. In Reference 197, the composing microservices is studied from the perspective of semantic annotations. This overlaps slightly with the second objective of our study since composing microservices entails reasoning about granularity. However, our work takes a broader approach to studying the aspects of reasoning about granularity. In Reference 198, the activities comprising the transition to microservices are inferred through a literature review (aligned to Objective 1 of our study). However, it does not focus on microservice granularity so our study is significant to understanding this microservitization challenge. In Reference 199, a broad approach is taken to study microservice literature. We acknowledge that it goes beyond our work in studying the research approaches utilised in the microservice community. Furthermore, this study dives into the computing aspect of the transition to microservices (analogous to the development publications category in our study). Nevertheless, our study is goes beyond199 with regards to defining the transition to microservices. Furthermore, our work is unique with regards to studying literature along different aspects of microservice granularity.

Several systematic literature reviews and surveys focus on defining the fundamental properties of microservices and the challenges of adopting them. They range in their rigour: some follow a rigorous search protocol12,18,45,73,177,200-202 while others are less formal.203 These studies overlap partially with Objective 1 of our study; they present activities related to microservices thereby contributing to a better understanding of the transition. However, they do not define on the transition to microservices nor can they be used to understand the microservice granularity problem. In Reference 175, the transition to microservices is partially described in terms of design patterns that can be applied to microservices. However, the scope of activities comprising the transition is not clearly defined. The approach of design patterns is also used in Reference 204. The patterns proposed in Reference 204, however, do not necessarily require microservice granularity adaptation. However, our work focuses on studying literature regarding microservice granularity. We envision that our work and References 175, 204 can benefit from each other to elicit granularity adaptation patterns and to expand the concerns of microservice architectural design patterns.

Some studies mainly focus on modeling microservices.205,206 Their focus partially overlaps with our following research question: what are the modeling approaches used to define the granularity of a microservice?

On the other hand,207 aims to “construct knowledge of quality attributes in architecture through a systematic literature review (SLR), (an) exploratory case study and (an) explanatory survey (p1 of Reference 207).” In essence, this study can be used to partially address our question: what are the quality attributes considered when reasoning about microservice granularity and how are they captured? Nevertheless, the other research questions of our study were not answered in this study.

Overall, the examined studies can complement this paper to discipline the understanding of the transition to microservices. In this paper, our research questions are formulated with focus on a specific problem of this transition—reasoning about microservice granularity.

7 | RELEVANT RESEARCH FIELDS

Although we focus on publications that have direct links to microservices in our study, we acknowledge that there are publications prior to the time period considered in this mapping study that could be relevant to the emergence of microservices and thereby to our research. Such publications do not fit our search strategy because they do not make direct reference to microservices. Nevertheless, in this section we briefly summarize the examined literature in these areas indicating how they can be relevant to our systematic mapping study objectives.

7.1 | Research fields relevant to understanding the microservice transition

Even among microservice adopters, there are still debates over distinguishing SOAs from microservice architectures.4,38,39,175,179,208-210 Therefore, we infer that contributions defining the properties and challenges of adopting SOA architectures can be relevant to understanding the transition to microservices.
Seminal work defines SOA as an architectural style which can guide business process definition\textsuperscript{211-216} and support “rapid, low-cost composition of distributed applications (p1 of References 217-219).” The SOA style was introduced to address architectural complexity, redundant programming and inconsistent interfaces.\textsuperscript{220,221} In SOAs a service is a self-describing unit that “consists of a contract, one or more interfaces and an implementation (p57 of Reference 222).” SOAs in turn comprise an application frontend, services, a service repository and enterprise service bus.

Despite the resemblance between microservices and SOAs, there is a subtle distinction we infer from our microservitization definition. The distinction comes from:\textsuperscript{3} (i) the potential of microservices as autonomous fine-grained computational units with lightweight communication mechanisms rather than service buses and, (ii) the operational and organizational flexibility enhanced by microservitization. Further elaboration of these points is presented in Reference 223.

### 7.2 Research fields relevant to understanding the microservice granularity problem

Modeling microservice granularity can be inspired by architectural modelling approaches—a wide research field, where contributions capture different notions of the architecture at varying levels of abstraction.\textsuperscript{224} Domain-driven modeling is the most relevant to the modeling approach we describe in Section 4.2 because it strives for logical isolation of business functionalities.\textsuperscript{29,225} However, domain-driven modeling is more concerned about the relationships between functional boundaries rather than their scope. In essence, domain-driven modeling can inform the structural rather than behavioral aspect of modeling microservices.

The Zachman framework\textsuperscript{226} provides a comprehensive guide to the different dimensions and perspectives for architectural modeling. Two dimensions in this framework are aligned with the modeling approach we call in Section 4.2: what the model units are and where these units are located relative to each other. We call for a modeling approach that explicitly captures what each microservice is concerned about and helps define where the business functionalities encapsulated by the microservice are located relative to each other.

A more dynamic architectural modeling approach is feature modeling,\textsuperscript{227} where an architecture is defined as a set of variability points, the candidates for each variability point and the rules constricting the dependencies across variability points. We appreciate this modeling technique is useful for formulating architectural decision-making problems. However, we would need to leverage such concept to focus on microservice boundaries being the variability point. While dependency rules in a feature model can give an insight about microservice granularity, they do not explicitly model it. ExecUtable RuntimEMegAmodels (EUREMA)\textsuperscript{228} provides a yet more powerful dynamic modeling technique. Here an adaptation engine contains a runtime model which represents the evolution of the system as well as adaptation activities to be executed on that model. The link between the adaptation activities and the runtime model is expressed using runtime mega-models. Similar to feature modelling, EUREMA needs to capture the notion of microservice boundaries more explicitly.

Boundaries are modelled more explicitly in design structure matrices (DSMs).\textsuperscript{229} Modularity metrics\textsuperscript{230} can be used to assess the degree of interdependence across these boundaries. We have not seen a dynamic application of DSMs that captures the changes in these dependencies across a time unit (eg, release cycles).

Since we call for objective reasoning about microservice granularity adaptation, design metrics can inspire this requirement since they provide an objective way to capturing attributes of a design decision. Effort-based metrics\textsuperscript{231} evaluate software development and maintenance efforts when a transition is made from centralised to distributed system architectures.\textsuperscript{232} By analogy, an objective way is needed to evaluate development and maintenance costs when microservice granularity is adapted. Metrics related to cohesion, coupling and visibility of system components are presented and visualized in References 232,233, which can be used to assess the impact of granularity adaptation on the microservice architecture's modularity.

Since reasoning about microservice granularity is in essence a dynamic architectural design decision problem, several software engineering fields can be relevant to it. Architectural analysis methods, architectural design patterns, service composition and orchestration approaches, runtime architectural adaptation, architectural refactoring and feedback control loops are only some of the relevant fields. In the following subsections we categorise contributions in these fields according to their level of autonomy:

- **Manual** contributions with full reliance on the software architect and/or stakeholders (eg, architectural design patterns)
**Partially autonomous** contributions where there is an autonomous agent but the software architect still makes the final decision regarding the optimal architecture (eg, interactive service orchestration and/or composition)

**Fully autonomous** contributions where the decision-making process and executing the decision are fully handled by an autonomous agent (eg, feedback control loops, online architectural refactoring)

### 7.2.1 Manual contributions

Out of the approaches in this subsection, the most cost- and value-aware approaches we examined are\textsuperscript{234,236} Refactoring the architecture according to patterns\textsuperscript{234} or to introduce modularity\textsuperscript{236} are regarded as value-bearing investments.\textsuperscript{235} However, these approaches are only applied statically at design time. In this paper, we call for a similar view for reasoning about microservice granularity.

In Reference 237, a cost benefit analysis method (CBAM) is proposed as a generic architecture evaluation method which utilizes techniques in decision analysis, optimization, and statistics to evaluate architectural design decisions. However, CBAM does not dynamically track and update the added value of architectural decisions. These dynamic updates are critical to the nature of the microservice granularity problem.

In Reference 238, the net benefit of a software is calculated by deducting its total costs from the total benefit. These are manually elicited and monetised from software architects through a series of questions (eg, “what is the status of the environment without the system?”). To our knowledge, this approach, however, does not consider the uncertainty in the answers to these questions nor does it update them at runtime. In Reference 239 the predictive analysis of design captures the value-driven impact of architectural decisions as multidimensional normalized, weighted cost, and value vectors. Therefore, it can only provide static objective decision support for reasoning about microservice granularity.

The techniques in Reference 240 present different architectural evaluation methods and their motivations. Software Architecture Analysis for Evolution and Reusability (SAAMER), Scenario-Based Architecture Re-engineering (SBAR) and Architecture Level Prediction of Software Maintenance (ALPSM) in particular take an objective approach to architectural evaluation.

SBAR captures the runtime nature of decision-making by providing different quality attribute evaluation techniques depending on whether the quality attribute is concerned with the “development” of the system (ie, design time, such as reusability, which is handled by scenario-based evaluation) or the “operation” of the system (ie, runtime, such as performance, which is handled by simulation-based evaluation). SAAMER on the other hand partially addresses granularity problem by analyzing the level of interaction between different scenarios of the system as a means of assessing the level of functionality isolation in the system. Nevertheless, both methods have not been explicitly applied in a dynamic environment to our knowledge.

ALPSM and CBAM take a value-driven approach to the evaluation which makes them more systematic architectural evaluation approaches than SAAMER and SBAR. Furthermore, ALPSM uses probabilities to capture the likelihood of the impact of scenarios and CBAM captures the uncertainty the architectural analysis. ALPSM and CBAM therefore partially capture uncertainty, although they do not operate at runtime and thereby they suffer from the limitation of design-time analysis.

Classical design patterns presented in Reference 241 extensively study creational (concerned with object instantiation), structural (concerned with relationships between objects) and behavioral patterns (concerned with coordination between objects). These patterns are further categorized according to their static or runtime nature. In that context, reasoning about microservice granularity can benefit from runtime creational design patterns. However, the design patterns of that category in Reference 241 do not capture the scope—boundary—where a pattern can be enforced. Service workflow patterns have been presented in a seminal work\textsuperscript{242} which implicitly discussed the issue of granularity is SOAs. However, we envision that the distinction between microservices and SOAs calls for explicitly addressing granularity adaptation decisions in the context of microservice constraints.

### 7.2.2 Partially autonomous contributions

In References 243,244, pattern-based engines are proposed to synthesize a composition of atomic and composite services. However, we envision that reasoning about microservice granularity needs to be grounded on objective rather than
pattern-based evaluation. In References 245,246, a microservice-specific approach for addressing the microservice granularity problem is proposed which relies on microservice web application log mining to extract the usage pattern and then making adaptive decisions regarding microservice granularity to ensure an economically sustainable architecture. We acknowledge this work is very closely related to the decision support called for in this paper. Nevertheless, it does not explicitly analyze the value-driven implications of such adaptive decisions as we call for in Section 4.2.

7.2.3 Fully autonomous contributions

The closest fully autonomous contribution to the effective decision support we call for in Section 4.2 is the ASTRO-CapEvo orchestration framework.247 It is a runtime framework that allows partial definition of business processes for service-based systems at design-time. Subsequently, the framework orchestrates “automatically composing the currently available services, provided by other actors and systems, according to the execution context and the goal of the process to be refined” using state transition systems. This framework takes a runtime approach to decision-making. However, the decision problem targeted by ASTRO-CapEvo is service composition rather than microservice granularity. Moreover, ASTRO-CapEvo does not consider objectively reason about composing the available services.

The Self-Serv framework presented in Reference 248 facilitates composite web service execution through peer-to-peer message exchange between coordination agents, which manage the service composition according to a static state chart. This framework can be utilised to address microservice granularity adaptation, but the knowledge that drives the service composition is static, meaning the runtime nature of the granularity problem is not captured in this framework.

Reputation-based dynamic service configuration techniques such as Reference 249 use a policy language to capture service consumers’ and providers’ profiles and then utilize these profiles to dynamically configure an optimal concrete service architecture. The work in Reference 250 leverages on the concept of reputation by capturing trust in the feedback given regarding the services. In particular, a model is proposed to aggregate feedback from several consumers of a service to reduce the effect of biased feedback. In both cases, a subjective user profile is used to drive the composition rather than an objective value- and cost-driven approach.

Case-based reasoning about concrete service composition is presented in Reference 251 where a solution space of composite services is formalized using recursive tuples of services. An agent then synthesises the optimal service decomposition given a request for services from the user which are then bound at runtime to concrete services fetched from a registry. A distinction is therefore made here between concrete service selection and the higher level composition of services; this can be utilised to address the granularity problem. However, this solution does not capture the dynamic nature of the microservice granularity problem.

Service composition techniques based on model checking252-254 dynamically adapt probabilistic models of the system according to runtime changes in the scenarios surrounding the system or runtime changes in the requirements of the system. In Reference 254 Bayesian learning improves service composition synthesis process through runtime knowledge updates about the system’s behavior over its lifetime. In Reference 253 an abstract service composition is mapped to a concrete service composition at runtime. The field of model-checking therefore is an attractive one for runtime decision-making. Such contributions however need to be leveraged for the specific problem we are concerned with (ie, the granularity of microservices).

Similar to the field of model checking is runtime architecture modeling. Several contributions in this field manifest runtime changes to the architecture.228,255-263 Other contributions are catered for systems which exhibit a similar level of dynamism to microservices.264-266 However, these contributions would need to be leveraged with objective reasoning that considers both the added value and cost of granularity adaptation.

Another approach of an autonomous solution is dynamic service formation rather than dynamic service composition. Frameworks such as References 267,268 provide means to dynamically produce web service specifications conforming to a service composition. These approaches can be utilized to complement the decision support we call for in this paper.

The runtime, uncertain context of microservice granularity adaptation calls for support similar to that provided by engineering self-adaptivity3,191 into an architecture. The role of a self-adaptive solution is to refine and update at runtime the architects’ design-time expectations about the architecture’s behavior. There are several mechanisms which can be adopted in this solution.269-273 Underlying most of them is the concept of feedback control loops274,275 which can be used “to Monitor, Analyse, Plan and Execute adaptations[16 of Reference 276]” in a system regarding trade-offs of concern (the phases of this loop are abbreviated as the MAPE loop). The knowledge learnt about the system needs to be maintained in a knowledge base (the phases of this loop are abbreviated as the MAPE-K271).
Control loops can be composed in a centralized, hierarchical, master-slave, or fully decentralized pattern. Each pattern varies in which components of the architecture carry out which phase(s) of the control loop. The centralized pattern is more suitable for monolithic architectures. In a hierarchical pattern, the full MAPE loop is effected at individual services, with higher level services having a more general view of the architecture. The individual service MAPE loops pass information to the higher level loops at short time intervals. Although this pattern is well-suited to addressing the trade-offs of concern here, its only shortcoming is deciding what the higher level component with the global view of the architecture should comprise and the possibility of this service creating a bottleneck. A variant of the hierarchical pattern is the master-slave pattern where the individual services only monitor and execute while a higher level service comprises the analysis and planning phases. Although more lightweight than the hierarchical pattern, the master/slave pattern suffers from the same shortcomings as the hierarchical pattern. The decentralized pattern on the other hand takes away the need for a service with a global view of the control loop. The MAPE loop is implemented in each service and information is passed across the services for decentralized management. The major challenge of this pattern is guaranteeing a consistent view of the system and its environment across all the control loops. However, this pattern is the most aligned with the autonomy of microservice architectures and the scale at which microservices operate.

Runtime architectural adaptation approaches have been proposed before in the Rainbow framework which is the most aligned with the concept of feedback loops. The Rainbow framework provides a reusable solution to induce self-adaptivity into a system in a cost-aware manner. However, it is debatable whether dynamically adapting the level of granularity of a microservice can be captured using the Rainbow framework. This is because the solution space here varies regarding the number of services used and the interaction patterns between them. To our knowledge, the Rainbow framework has not been applied to such a setting before. Another prominent approach is the architectural refactoring approach where a set of anti-patterns is proposed which can be detected dynamically and used to trigger refactoring an architecture. This approach has as its main motive enhancing the modularity of the architecture rather than reasoning about modularity in an objective manner.

Realising microservice granularity adaptation involves changes to a deployed architecture. These activities are similar to those underlying the field of online architectural refactoring. Several contributions in this field pave the way to automated online architectural refactoring and modeling transformations. These contributions are driven by meeting a specific architectural design pattern, but they do not objectively reason about granularity adaptation.

In the field of AI planning, an ontology-based approach to architectural adaptation is proposed. A shared ontology of generic “procedures” (or templates) is produced which the stakeholder can choose from at run-time. An agent then executes this procedure customizing it depending on the scenario in which the system will operate. It is appreciated that the use of ontologies promotes sharing knowledge across across architects. However, we envision that the decision support we are calling for in this paper can promote knowledge sharing and profiling to inform reasoning about microservice granularity.

8 | CONCLUSION

In this paper we report on a systematic mapping study to consolidate various views, principles, methods and techniques that are commonly adopted to assist the transition to microservices. We systematically describe the study’s process and report its results. We contribute a working definition capturing the fundamentals of the transition; we term it as microservitization. Microservitization is a form of servitization where services/components are transformed into microservices — a more fine-grained and autonomic form of services—to introduce added value to the architecture. Microservitization is also an example of a paradigm shift since it involves a dramatic change to the way technical activities are carried out and aligned with a microservice adopter’s business objectives. We then shed light on a fundamental problem of microservitization: microservice granularity and reasoning about its adaptation as first-class entities. This study has reviewed and identified gaps in the state-of-the-art and -practice that relate to the modeling approaches, aspects considered, guidelines, and processes used to reason about microservice granularity. The identified gaps pave the way to opportunities for future research and development related to reasoning about microservice granularity.

In particular, we identify there is room for future research regarding:

- A systematic architecture-oriented modeling support for microservice granularity that facilitates runtime analysis of microservice granularity in a systematic architecture-oriented manner.
A dynamic architectural evaluation approach to reason about the cost and added value of granularity adaptation.

Moreover, the findings of our study can be further disciplined in the future both horizontally and vertically. Horizontal disciplining entails including/excluding more microservitization activities (e.g., securing communication links, cloud resource scheduling and data management) and more quality attributes related to reasoning about granularity (e.g., security). Vertical disciplining entails using other evidence-based approaches which extract evidence from multiple sources (e.g., interviewing microservice practitioners, industrial case studies).

ORCID
Sara Hassan https://orcid.org/0000-0001-7481-0434

REFERENCES
6. Newman S. Practical considerations for microservice architectures. Part 2. (Sam Newman, UK); 2014. https://www.youtube.com/watch?v=cU0J0w6sFoA.


32. Krause L. Microservices: patterns and applications designing fine-grained services by applying patterns; 2015. https://books.google.co.uk/books?id=dd5-rgEACAAJ.


34. Miles R. Antifragile software building adaptable software with microservices. Leanpub; 2016.


