Comparison of Tools for Software Architecture Extraction of Asynchronous Microservice Systems

Jonas Frey

Institute of Information Security and Dependability (KASTEL)
Advisor: M.Sc. Snigdha Singh

English abstract.

Keywords

1 Introduction

This chapter will provide a motivation of why a systematic literature review about tools for architecture extraction of asynchronous systems is necessary.

In chapter 2, we will provide some foundation knowledge. In chapter 3, we will talk about the design and goal of this literature review and the selection of the papers. In chapter 4, we will present the results and compare them using five different aspects. We will discuss these results in chapter 5 and talk about related work in chapter 6, before drawing conclusions in chapter 7.

2 Foundation

In the following sections, we will provide some foundation knowledge for the rest of the paper.

2.1 Microservice Architecture

In a microservice architecture, the software is partitioned into many small components ("microservices"), which operate independently of each other and communicate via messages [Dra+17]. This software architecture style allows the construction of highly reusable components which focus on a single task (e.g. applying a watermark to a video). This loose coupling allows for independent teams to work on different components or even the use of off-the-shelf components. Also, scaling the application can be achieved by

simply duplicating the bottlenecked microservices [Dra+17]. A microservice is defined by its provided- and required-interfaces, which allow for the message-exchange with other components [SKK21].

2.2 Asynchronous RESTful Communication

Microservices that communicate asynchronously are typically realized in one of two kinds. Either using a RESTful pattern or using message-based communication.

Some systems use microservices that communicate asynchronously via asynchronous HTTP REST interfaces. There are two interaction scenarios for this kind of asynchronous communication. One possibility is that the initial HTTP request returns an HTTP code 202 (Accepted) and provides a location where the microservice can query the status of the operation. Once the operation on the server has finished, the provided location will return the results of the query. [MW18]

Alternatively, the microservice may be required to provide a callback method (e.g. a web hook, [Lin]) where the server can send the results once the operation has finished [MW18].

We will refer to both of these communication methods as RESTful asynchronous communication in the following paper.

2.3 Message-based Communication

Contrary to RESTful asynchronous communication, other microservice systems use message-based asynchronous communication. These systems deploy a message broker, a component that other components dynamically bind to. This message broker is then used to send and receive messages and is responsible for the distribution of these messages. Using a message broker allows for better system performance. [SKK21]

2.4 Software Architecture Extraction (SAR)

Software Architecture Extraction (SAR) is the process of reverse engineering a software architecture from a given system. Vital for building this architecture is information about the communication between the different components. This information is extracted using either a static (using only static inputs, e.g. source code), dynamic (using runtime information, e.g. logs) or hybrid (using both) approach. In the context of asynchronous communication, a static extraction algorithm would in the case of a RESTful asynchronous communication, analyze the HTTP calls made in code to determine the relationships between the components. In the case of message-based communication, a static approach is unable to extract a useful architecture, since message-based systems exchange those messages only at runtime and thus the required information about which components communicate with each other can only be retrieved as part of a dynamic or hybrid analysis. [SKK21; MW18]

2.5 Palladio Component Model

The Palladio Component Model (PCM) is a meta-model for the description of component-based software [BKR09]. It is used to predict the performance properties of component-based software at design-time by specifying a model of the system, its components and how the system is going to be used. ***[TODO: extend]***

3 Study Design

This chapter will explain the design of the systematic literature review and how it was executed.

3.1 Study Aim

The aim of this paper is to find and compare the tools available for the extraction of the architecture of asynchronous microservice systems. For this purpose, we define two research questions.

[TODO: combine 3.1 and 3.2 into one section?]

3.2 Research Questions

The two research questions we want to answer in this paper are

RQ1. What are the tools available for the extraction of asynchronous architectures of microservice systems?

RQ2. To what extend do the tools support software architecture extraction?

3.3 Selecting the Papers

The search for research papers was performed by performing several queries using Google Scholar. The following queries were used to search for papers:

- `architecture (extraction OR reconstruction) (dynamic OR logs OR asynchronous) microservice`
- `("architecture extraction" OR "architecture reconstruction") (dynamic OR logs OR asynchronous) microservice`
- `reverse engineering (dynamic OR logs OR asynchronous) (microservice OR mixed-technology)`

Additionally, the references of the found results were used to look for further papers. For a paper to be selected, it had to match our selection criteria depicted in Table 1

In total, we will look at five papers, which each present an approach for the extraction of asynchronous architecture of microservice systems.

The papers are

1. ARCHI4MOM [SWK22], [SKK21]

Inclusion	Papers that present an approach for extracting microservice architectures Papers that are able to extract asynchronous architectures
Exclusion	Papers that only present a foundation or compare other approaches

Table 1: Inclusion and exclusion criteria for selecting the papers

2. MiSAR [AAE18]
3. — [BHK11]
4. MICROLYZE [Kle+18]
5. — [MW18]

4 Results

Table 2 shows the results of the comparison in tabular format. In the following sections, we will talk about each paper individually.

Each approach will be presented using five aspects:

1. **Input** (e.g. source code or logs)
2. **Approach** (how the extraction process works)
3. **Output** (e.g. PCM or UML)
4. **End user** (who the result is intended for)
5. **Evaluation** (how were the results evaluated)

4.1 ARCHI4MOM

Input. The ARCHI4MOM approach extends the Performance Model Extraction (PMX) approach [Wal+17; SWK22] and therefore takes the same inputs. The first step in the ARCHI4MOM approach is to instrument the source code with the Jaeger tracing tool [1] to collect tracing data. Using the OpenTracing API, ARCHI4MOM then instruments all microservices to generate trace data, which can be used later to reconstruct the asynchronous communication between the microservices. [SWK22]

Approach. ARCHI4MOM extends the Performance Model Extraction (PMX) approach [Wal+17; SWK22] to support asynchronous communication. This is achieved by adding a dependency to the OpenTracing API to each microservice to introduce a new set of information that was not present earlier [SWK22]. This tracing data is then collected in

[1] https://www.jaegertracing.io/

the form of JavaScript Object Notation (JSON) files, which become the input of the next phase [SWK22].

The JSON files will then be used to analyze the structure of the traces. For message-based asynchronous communication, this presents a challenge since the information is distributed over different *spans* whereas using synchronous communication, it would be in a single span [SWK22]. To match this inter-component communication using middleware, the called method needs to be matched using the OpenTracing API tags [SWK22]. In the next step, the ARCHI4MOM approach looks for send operations that do not have information about the topic ***[TODO: explain]*** they send to and fills in this information by retrieving it from other send operations in the tracing data [SWK22]. Then the approach iterates over all sending spans that have a FOLLOWS-FROM or *message-bus* relation ***[TODO: explain]*** tag and propagate their topics to the receiving spans [SWK22].

To reconstruct message-based communication using PMX, the authors extended PMX to support the required Palladio Component Model elements [2]. The authors then implement additional PMX logic to be able to reconstruct asynchronous architectures using these model elements. [SWK22].

Output. The output of the ARCHI4MOM approach is a Palladio Component Model (PCM) [SWK22]. This PCM contains the extracted components, as well as the interfaces for communication between each other [SWK22]. The communication channels are represented by *DataChannel*s (representing the middleware) and *DataInterface*s (representing the type of data the interface can send/receive), which are created in the PCM repository as part of the extraction [SWK22].

End User. The output of the ARCHI4MOM approach is a PCM. This model can then be used by software architects together with usage scenarios to create simulations, predicting the non-functional properties of the software.

Evaluation. To evaluate the approach, the authors created a manual PCM of the Flowing Retail sample application [3]. This manual model was then verified by three developers to be correct and compared to the automatically extracted model using a Goal Question Metric (GQM) plan [Van+02; SWK22]. Using this plan, both sets of model elements (manual and automatic) are then compared using Precision, Recall and F1 score [SWK22]. The automatic approach achieved a precision score of 100%, a recall score of 95.65% and an F1 score of 97.8% [SWK22]. ***[TODO: should I include the results or not?]***

4.2 MiSAR

Input. MiSAR extracts and gathers different data from static artifacts. This data includes docker files that assemble the containers for the microservices, docker compose files that orchestrate multi-docker-container systems, java source code, maven pom.xml files, YAML configuration files, documentation and tool support [AAE18]. The java source code is

[2] https://github.com/PalladioSimulator/Palladio-Addons-Indirections/tree/master/bundles/org.palladiosimulator.indirections/model

[3] https://github.com/berndruecker/flowing-retail/tree/master/kafka/java

reverse engineered using a tool called Enterprise architect [4], providing UML class diagrams from the source code [AAE18]. Additionally, Zipkin [5] is used to trace communication between microservices to build a call graph [AAE18]. Information about latencies was retrieved using TCPDump [6] and information about the ports, IP addresses of container, and connectivity between containers were extracted using the Sysdig tool [7] [AAE18]. This information is all stored in a repository for further use.

Approach. MiSAR is a manual approach that is executed in two phases [AAE18]. The first phase (Recovery Design, RD) defines architectural concepts, which are extracted in the second phase (Recovery Execution, RE) [AAE18]. ***[TODO: extend]***

Output.

End User.

Evaluation.

4.3 — [BHK11]

Input. The approach by Brosig et al. uses onlyl monitoring data collected at runtime to extract the effective architecture [IWF07] of the system.

Approach. The approach presented by Brosig et al. only extracts the effective architecture [IWF07] of the system, meaning that only parts that are effectively used at runtime are considered [BHK11]. The first step in the extraction of the effective architecture. Before the extraction process can begin, component boundaries, which separate components as single entities from the point of view of the system's architect, need to be defined [BHK11]. This can be either done manually by a software architect or automatically using static code analysis [BHK11]. After the system boundaries have been determined, the running system is monitored using *call path tracing* [BHK11]. The resulting *event records* (representing entries or exists of components) are grouped together into *call path event record sets* which contain event records that were triggered by the same system request [BHK11]. This data can then be used to obtain a call path [BHK11] as well as a list of external services, a component's provided interface calls. To extract an accurate representation of the system's components and connections between them, a representative usage profile has to be chosen, as the extraction only captures the actual communication that happens during the extraction approach [BHK11]. After the components and their connection have been extracted, the component-internal performance-relevant control flow has to be modeled. This includes the internal behavior as well as the external service calls, the component makes [BHK11].

[4] http://www.sparxsystems.com.au/products/ea/
[5] https://zipkin.io
[6] https://www.tcpdump.org
[7] https://github.com/draios/sysdig

For the second step, the approach aims to extract model parameters for performance prediction [BHK11]. This is achieved by extracting branch probabilities and loop iteration numbers from the call paths [BHK11]. For branching probabilities, mean values are used, whereas loop iteration numbers are represented by a Probability Mass Function (PMF) to allow for accurate representation of cases where a loop is for example either executed twice or ten times [BHK11]. Next, the approach tries to quantify the resource demands (e.g. CPU, HDD) of the components' internal computations [BHK11]. This demand is represented by the total processing time minus the time spent waiting for the resource to become available [BHK11]. These parameters are then averaged over the observed call paths [BHK11]. The approach is also able to handle e.g. branches that depend on input parameter values [BHK11]. In this case, the dependency has to be known a-priori for the approach to quantify these dependencies [BHK11].

In the third step, the performance model is calibrated by comparing its predictions with measurements on the real system [BHK11]. The correction to be done when measuring a deviation between the prediction and the measurement is done by increasing a factor of overhead and accounting for delays produced by the middleware stack, the system runs on [BHK11].

Output. The output of the approach by Brosig et al. is a performance model [BHK11]. In their proof-of-concept implementation, they generate a Palladio Component Model (PCM) [BHK11].

End User. The end user for this approach is e.g. a software architect, which uses the performance model to predict software quality attributes.

Evaluation. The evaluation of the approach was accomplished by implementing it and applying it to a case study of a Java Enterprise Edition application [BHK11]. The application used was the SPECjEnterprise2010 benchmark [8].

4.4 MICROLYZE [Kle+18]

Input. Microlyze uses runtime data from a service discovery service, as well as manual inputs (e.g. semantic descriptions, mappings to technical requests) to reconstruct the software architecture [Kle+18].

Approach. The Microlyze recovery approach is executed in six phases, which are meant to be executed continuously [Kle+18]. The first phase rebuilds the current system architecture by checking a service discovery service (e.g., Eureka [9] or Consul [10]) and updating the status of the services in the current architecture [Kle+18].

The second phase uses the IP addresses and ports retrieved from the service discovery service to establish a link between the services and the hardware used [Kle+18]. Together with the IP addresses and ports, a monitoring agent has to be installed on each hardware

[8] https://www.spec.org/jEnterprise2010/
[9] https://github.com/Netflix/eureka
[10] https://www.consul.io

component to retrieve additional information about the hardware [Kle+18]. Additionally, a monitoring probe is installed on each microservice to observe the HTTP communication [Kle+18]. The probe injects tracing data in the HTTP headers and therefore helps to gather timing data [Kle+18]. The approach then collects additional infrastructure and software-specific data (e.g., endpoint name, class, method, HTTP request) and attaches this information as annotations [Kle+18]. Using this data, the approach is then able to detect the dependencies between the microservices by following identifiers in the HTTP requests during runtime [Kle+18]. This tracing is done using zipkin [11], streamed via apache kafka to Microlyze and then stored in a cassandra database [Kle+18]. Microlyze additionally classifies each service on basis of the distributed tracing data [Kle+18]. For example, if the first accessed microservice is identical in most requests, it is classified as a gateway service [Kle+18].

In the third phase, all user transactions are stored in a database, including what the user does and in which order [Kle+18]. These user transactions are then mapped to business transactions using a business process modeller [Kle+18]. The information about the user transactions is used to create an association between the business transactions and the microservices that process these transactions [Kle+18].

The fourth phase is responsible for defining semantic descriptions to each business activity, that can be performed by a user (e.g. *register*, *open shopping cart*, etc.) [Kle+18].

After the business activities have been augmented with descriptions, phase five is able to create a mapping between the business activities ("a sequence of related events that together contribute to serve a user request" [Kle+18]) and the technical requests extracted by zipkin [Kle+18]. For this purpose, the authors enhanced the business process modeller with the regular expression language in order to describe technical requests [Kle+18]. These regular expressions are stored in the database and matched on new incoming transactions to detect, if they might refer to an already modelled business activity [Kle+18].

Finally, the sixth phase polls the service discovery service continuously to receive updates about unregistered or newly registered services [Kle+18]. Services that are no longer registered are marked as such and unknown user requests that cannot be mapped to an existing business activity using the regular expressions are added to a list of unmapped URL endpoints [Kle+18]. Changes to the underlying infrastructure are detected by changes in IP addresses or ports and lead to automatic adaptions in the architecture model [Kle+18].

Output. The results of the architecture extraction are displayed to the user as an adjacency matrix [Kle+18].

End User. The end users of the approach are administrators and enterprise or software architects [Kle+18].

Evaluation. To evaluate the appraoch, the authors developed a prototype and applied it to the TUM LLCM platform [12] [Kle+18]. They developed a service called *Travelcompan-*

[11] https://github.com/openzipkin/zipkin

[12] https://www.cs.cit.tum.de/bpm/krcmar/research/finished-projects/tum-llcm-tum-living-lab-connected-mobility/

ion which is meant to form travel groups to save on travel costs [Kle+18]. To discover the relationships between the components, the authors produced traffic using JMeter [13] [Kle+18]. After step three completed, the authors enhanced the technical transactions with a business semantic and mapped the business activities to user transactions [Kle+18].

4.5 — [MW18]

Input. The approach uses static information, provided by configuration files and static information about services (e.g., name, version, etc.) [MW18]. Additionally, runtime communication between the microservices is collected and aggregated [MW18].

Approach. The approach presented by Mayer and Weinreich works in three steps, which are modeled by three main components [MW18]. The first step—data collection—uses the *Data Collection Library* to collect and provide static and runtime data for the architecture extraction components [MW18]. This library collects and provides service-, interaction-, and infrastructure-related information using Swagger [14] to generate API descriptions and infrastructure-specific information providers (e.g., configuration files) that are configured by the user [MW18]. The second step—data aggregation—uses the *Aggregation Service* to aggregate the information collected in the first step [MW18]. Lastly, the third step—data combination—uses the *Management Service* combines the information from the different microservices and stores it in a data model [MW18].

The authors also differentiate between three different kinds, or phases, of architecture extractions [MW18]. The first phase is the static information extraction. This extraction starts after a new service is deployed [MW18]. Using swagger, a JSON representation of the service is sent to the *Management Service*, which uses this information (namely the name and version of the service) to identify, whether the deployed service is a new instance of an already existing service, or a new service altogether [MW18]. This phase of static architecture extraction finishes, by storing all static service information in the central *Management Service* information database [MW18].

The second phase concerns the extraction of infrastructure information [MW18]. This phase builds on the information collected in the first phase and creates the according service instance node in the *Management Service*'s database [MW18]. This database is stored as a graph with directed edges, connecting the newly inserted service instance node to the necessary services and physical infrastructure information [MW18]. If the nodes representing the physical infrastructure the service was deployed on (host and region) do not already exist, the *Management Service* creates them. As with the first phase, all information is stored in the information database at the end of the extraction [MW18].

The third and last phase is responsible for extracting the runtime information. For this purpose, all outgoing and incoming requests of a microservice are logged to a local file, including their timestamp, response time, response code, the ID of the source service instance, the URL of the target service instance, and the requested method [MW18]. The *Aggregation Service* consumes these log files periodically via REST interfaces and aggre-

[13] https://jmeter.apache.org
[14] https://swagger.io

gates them [MW18]. The aggregation condenses requests that have the same source and destination instance, the same method and the same response [MW18]. The aggregated requests contain the time interval, the number of requests, and the average, maximum, and minimum response time [MW18]. Lastly, these aggregated requests are then sent to the *Management Service*, which stores them in its database [MW18]. The *Management Service* can also use this information to mark services as inactive, if it receives no runtime information anymore [MW18].

Output. The output of this approach is a database containing a directed graph that represents the system's architecture, as well as aggregated runtime information (e.g., response times) about the different requests [MW18].

End User. The end users of the output of this approach are architects, developers, and operation experts [MW18].

Evaluation. The authors conducted a combined survey and interview study ([MW17]) and used the feedback to construct a microservice dashboard that supports the different use cases identified by the survey [MW18]. The authors then built a test scenario consisting of three microservices to test the long-term data collection capabilities of their approach [MW18].

5 Discussion

This chapter will discuss the results of the previous chapter.

6 Related Work

This chapter presents other papers which are similar to my work. For example [DP09], which compares different SAR approaches to formulate a state-of-the-art approach or [GIM13], which compares different SAR tools. We will also talk about the fact that [Gra+17] and [Lan+16] could be extended to support asynchronous communication in the future.

7 Conclusion

In this chapter, we will recap the findings that we made and finish the paper with concluding remarks.

References

[AAE18] Nuha Alshuqayran, Nour Ali, and Roger Evans. "Towards Micro Service Architecture Recovery: An Empirical Study". In: 2018. DOI: 10.1109/ICSA.2018.00014.

Name	Input	Approach	Output	End User	Evaluation	Year	Type
ARCHI4MOM ([SWK22])	source code	Extend PMX to support asynchronous architectures	PCM	software architect	Comparison with manual architecture	2022	tool
MiSAR ([AAE18])	source code, descriptive files, runtime traces	manual extraction approach ***[TODO: extend]***				2018	manual approach
— ([BHK11])	run-time monitoring data	combine an existing call path tracing and resource demand estimation to an end-to-end model extraction	PCM		SPECjEnterprise2010 benchmark application; comparison of prediction with measurements	2011	tool
MICROLYZE ([Kle+18])	monitoring data	continuously monitor for system changes using a service discovery service; trace HTTP requests using zipkin	database with services and their relations; web application to visualize as adjacency matrix	system administrators and enterprise or software architects	approach was applied to TUM LLCM platform, Travelcompanion service; traffic was generated and result was manually checked	2018	tool
— ([MW18])	static service information, infrastructure information and runtime logs	condense static and dynamic information into single dimension to analyze the evolution over time; visualize information in dashboard	aggregated data, visualized in dashboard		use tool in testing environment; check viability of results	2018	tool

Table 2: Results

[BHK11] Fabian Brosig, Nikolaus Huber, and Samuel Kounev. "Automated extraction of architecture-level performance models of distributed component-based systems". In: IEEE, Nov. 2011, pp. 183–192. ISBN: 978-1-4577-1639-3. DOI: 10.1109/ASE.2011.6100052.

[BKR09] Steffen Becker, Heiko Koziolek, and Ralf Reussner. "The Palladio Component Model for Model-driven Performance Prediction". In: *Journal of Systems and Software* 82 (2009), pp. 3–22. DOI: 10.1016/j.jss.2008.03.066. URL: http://dx.doi.org/10.1016/j.jss.2008.03.066.

[DP09] S. Ducasse and D. Pollet. "Software Architecture Reconstruction: A Process-Oriented Taxonomy". In: *IEEE Transactions on Software Engineering* 35 (4 July 2009), pp. 573–591. ISSN: 0098-5589. DOI: 10.1109/TSE.2009.19.

[Dra+17] Nicola Dragoni et al. *Microservices: Yesterday, Today, and Tomorrow*. 2017. DOI: 10.1007/978-3-319-67425-4_12.

[GIM13] Joshua Garcia, Igor Ivkovic, and Nenad Medvidovic. "A comparative analysis of software architecture recovery techniques". In: *2013 28th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE. 2013, pp. 486–496.

[Gra+17] Giona Granchelli et al. "Towards recovering the software architecture of microservice-based systems". In: 2017. DOI: 10.1109/ICSAW.2017.48.

[IWF07] Tauseef Israr, Murray Woodside, and Greg Franks. "Interaction tree algorithms to extract effective architecture and layered performance models from traces". In: *Journal of Systems and Software* 80.4 (2007), pp. 474–492.

[Kle+18] Martin Kleehaus et al. "MICROLYZE: A framework for recovering the software architecture in microservice-based environments". In: vol. 317. 2018. DOI: 10.1007/978-3-319-92901-9_14.

[Lan+16] Michael Langhammer et al. "Automated extraction of rich software models from limited system information". In: 2016. DOI: 10.1109/WICSA.2016.35.

[Lin] J Lindsay. *Web hooks to revolutionize the web (2007)*. Tech. rep. URL: https://web.archive.org/web/20180828032936/http://progrium.com/blog/2007/05/03/web-hooks-to-revolutionize-the-web/.

[MW17] Benjamin Mayer and Rainer Weinreich. "A dashboard for microservice monitoring and management". In: *2017 IEEE International Conference on Software Architecture Workshops (ICSAW)*. IEEE. 2017, pp. 66–69.

[MW18] Benjamin Mayer and Rainer Weinreich. "An Approach to Extract the Architecture of Microservice-Based Software Systems". In: 2018. DOI: 10.1109/SOSE.2018.00012.

[SKK21] Snigdha Singh, Yves Richard Kirschner, and Anne Koziolek. "Towards extraction of message-based communication in mixed-technology architectures for performance model". In: 2021. DOI: 10.1145/3447545.3451201.

[SWK22] Snigdha Singh, Dominik Werle, and Anne Koziolek. “ARCHI4MOM: Using Tracing Information to Extract the Architecture of Microservice-Based Systems from Message-Oriented Middleware”. In: *European Conference on Software Architecture* (2022).

[Van+02] Rini Van Solingen et al. “Goal question metric (gqm) approach”. In: *Encyclopedia of software engineering* (2002).

[Wal+17] Jürgen Walter et al. “An Expandable Extraction Framework for Architectural Performance Models”. In: *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion*. ICPE ’17 Companion. L’Aquila, Italy: Association for Computing Machinery, 2017, pp. 165–170. ISBN: 9781450348997. DOI: 10.1145/3053600.3053634. URL: https://doi.org/10.1145/3053600.3053634.